UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

ALGORITHM DESIGN FOR SCHEDULING AND MEDIUM ACCESS CONTROL IN HETEROGENEOUS MOBILE NETWORKS

Autor: Gek Hong Sim, IMDEA Networks Institute, University Carlos III of Madrid
Director: Joerg Widmer, IMDEA Networks Institute
Tutor: Ruben Cuevas Rumin, University Carlos III of Madrid

DEPARTAMENTO DE INGENIERÍA TELEMÁTICA

Leganés (Madrid), marzo de 2016

UNIVERSIDAD CARLOS III DE MADRID

PH.D. THESIS

ALGORITHM DESIGN FOR SCHEDULING AND MEDIUM ACCESS CONTROL IN HETEROGENEOUS MOBILE NETWORKS

Author: Gek Hong Sim, IMDEA Networks Institute, University Carlos III of Madrid
Director: Joerg Widmer, IMDEA Networks Institute
Tutor: Ruben Cuevas Rumin, University Carlos III of Madrid

DEPARTMENT OF TELEMATIC ENGINEERING

Leganés (Madrid), March 2016

Algorithm Design for Scheduling and Medium Access Control in heterogeneous mobile networks

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Prepared by
Gek Hong Sim, IMDEA Networks Institute, University Carlos III of Madrid

Under the advice of
Joerg Widmer, IMDEA Networks Institute
Ruben Cuevas Rumin, University Carlos III of Madrid
Departamento de Ingeniería Telemática, Universidad Carlos III de Madrid

---

Date: marzo, 2016

Web/contact: allyson.sim@imdea.org

This work has been supported by IMDEA Networks Institute.

TESIS DOCTORAL

ALGORITHM DESIGN FOR SCHEDULING AND MEDIUM ACCESS CONTROL IN HETEROGENEOUS MOBILE NETWORKS

Autor: Gek Hong Sim, IMDEA Networks Institute, University Carlos III of Madrid
Director: Joerg Widmer, IMDEA Networks Institute
Tutor: Ruben Cuevas Rumin, University Carlos III of Madrid

Firma del tribunal calificador:

Presidente:

Vocal:

Secretario:

Calificación:

Leganés, de de

Acknowledgements

In the success of completing this thesis and my PhD studies, many deserved my appreciation. First and foremost, I would like to deliver my utmost appreciation to my advisor, Dr. Joerg Widmer. Throughout the period of my PhD studies, he has been very patience and provided continuous encouragement, support, and guidance.

Secondly, Dr. Balaji Rengarajan, although we were only working together for a short one year period, he has been very helpful in the discussion of research ideas. Special thanks to Dr. David Malone and Dr. Cristina Cano for hosting me during my internship in Trinity College and working together to produce a good research paper. Thanks to Dr. Paul Patras and Rui Li from University of Edinburg for agreeing to collaborate with me and always offer to help in any way they could. I would also like to convey my special gratitude to the following people whom I met throughout my PhD studies and have been there to make IMDEA Networks and Trinity colleage a fun place to be in: Qing Wang, Dr. Thomas Nitsche, Dr. Ignacio De Castro, Dr. Adrian Loch, Dr. Bahar Partov, Dr. Jordi Arjona, Evgenia, Cristian Vitale, Elli, Hany Assasa, Camilo, Foivos, Roderick, and Aymen. Many thanks to my friends, Lydia Rajendran, Tava Manggai, Amin Ghazanfari, Nastaran Meftahi, and Suhanya Jayaprakasam, for always being there and care for my progress.

My great appreciation also goes to my parents and siblings for the unconditional love and support. Last but not least, an exceptional gratitude goes to my fiancé, Dr. Arash Asadi, for the unconditional support, continuous motivation, love and care as well as for being understanding throughout these tough years.

Abstract

The rapid growth of wireless mobile devices has led to saturation and congestion of wireless channels – a well-known fact. In the recent years, this issue is further exacerbated by the ever-increasing demand for traffic intensed multimedia content applications, which include but are not limited to social media, news and video streaming applications. Therefore the development of highly efficient content distribution technologies is of utmost importance, specifically to cope with the scarcity and the high cost of wireless resources. To this aim, this thesis investigates the challenges and the considerations required to design efficient techniques to improve the performance of wireless networks. Since wireless signals are prone to fluctuations and mobile users are, with high likelihood, have difference channel qualities, we particularly focus on the scenarios with heterogeneous user distribution. Further, this dissertation considers two main techniques to cope with mobile users demand and the limitation of wireless resources. Firstly, we propose an opportunistic multicast scheduling to efficiently distribute or disseminate data to all users with low delay. Secondly, we exploit the Millimeter-Wave (mm-Wave) frequency band that has a high potential of meeting the high bandwidth demand. In particular, we propose a channel access mechanism and a scheduling algorithm that take into account the limitation of the high frequency band (i.e., high path loss).

Multicast scheduling has emerged as one of the most promising techniques for multicast applications when multiple users require the same content from the base station. Unlike a unicast scheduler which sequentially serves the individual users, a multicast scheduler efficiently utilizes the wireless resources by simultaneously transmitting to multiple users. Precisely, it multiplies the gain in terms of the system throughput compared to unicast transmissions. In spite of the fact that multicast schedulers are more efficient than unicast schedulers, scheduling for multicast transmission is a challenging task. In particular, base station can only chose one rate to transmit to all users. While determining the rate for users with a similar instantaneous channel quality is straight forward, it is non-trivial when users have different instantaneous channel qualities, i.e., when the channel is heterogeneous. In such a scenario, on one hand, transmitting at a low rate results in low throughput. On the other hand, transmitting at a high rate causes some users to fail to receive the transmitted packet while others successfully receive it but with a rate lower than their maximum rate. The most common and simplest multicasting technique, i.e., broadcasting, transmits to all receivers using the maximum rate that is supported by the worst receiver.

In recent years, opportunistic schedulers have been considered for multicasting. Opportunistic multicast schedulers maximize instantaneous throughput and transmit at a higher rate to serve only a subset of the multicast users. While broadcasting suffers from high delay for all users due to low transmission rate, the latter causes a long delay for the users with worse channel quality as they always favor users with better channel quality. To address these problems, we designed an opportunistic multicast scheduling mechanism that aims to achieve high throughput as well as low delay. Precisely, we are solving the finite horizon problem for multicasting. Our goal is that all multicast users receive the same amount of data within the shortest amount of time.

Although our proposed opportunistic multicast scheduling mechanism improves the system throughput and reduces delay, a common problem in multicast scheduling is that its throughput performance is limited by the worst user in the system. To overcome this problem, transmit beamforming can be used to adjust antenna gains to the different receivers. This allows improving the SNR of the receiver with the worst channel SNR at the expense of worsening the SNR of the better channel receivers. In the first part of this thesis, two different versions of the finite horizon problem are considered: (i) opportunistic multicast scheduling and (ii) opportunistic multicast beamforming.

In recent years, many researchers venture into the potential of communication over mm-Wave band as it potentially solves the existing network capacity problem. Since beamforming is capable to concentrate the transmit energy in the direction of interest, this technique is particularly beneficial to improve signal quality of the highly attenuated mm-Wave signal. Although directional beamforming in mm-Wave offers multi-gigabit-per-second data rates, directional communication severely deteriorates the channel sensing capability of a user. For instance, when a user is not within the transmission coverage or range of the communicating users, it is unable to identify the state of the channel (i.e., busy or free). As a result, this leads to a problem commonly known as the *deafness* problem. This calls for rethinking of the legacy medium access control and scheduling mechanisms for mm-Wave communication. Further, without omni-directional transmission, disseminating or broadcasting global information also becomes complex. To cope with these issues, we propose two techniques in the second part of this thesis. First, leveraging that recent mobile devices have multiple wireless interface, we present a dual-band solution. This solution exploits the omni-directional capable lower frequency bands (i.e., 2.4 and 5 GHz) to transmit control messages and the mm-Wave band for high speed data transmission. Second, we develop a decentralized scheduling technique which copes with the deafness problem in mm-Wave through a learning mechanism.

In a nutshell, this thesis explores solutions which (i) improve the utilization of the network resources through multicasting and (ii) meet the mobile user demand with the abundant channel resources available at high frequency bands.

Table of Contents

List of Tables

List of Figures

List of Acronyms

CSMA/CA Carrier Sense Multiple Access with Collision Avoidance

mm-Wave Millimeter-Wave

IPTV Internet Protocol Television

LTE Long Term Evolution

eMBMS Evolved Multimedia Broadcast Multicast Service

BS Base Station

OMS Opportunistic Multicast Scheduling

MCS Modulation and Coding Scheme

LDPC Low-Density Parity Check

LT Luby Transform

GOP Groups of Pictures

Dyn-Prog Dynamic Programming

Weighted-CT Weighted Completion Time

LCG Least Channel Gain

SNR Signal-to-Noise Ratio

CSI Channel State Information

PER Packet Error Rate

PSR Packet Success Rate

Weighted-CTe Weighted-CT with rate estimation

TTI Transmission Time Interval

QPSK Quadrature Phase Shift Keying

QAM Quadrature Amplitude Keying

CQI Channel Quality Indicator

DVD Digital Versatile Disc

FH-OMB Finite Horizon-Opportunistic Multicast Beamforming

MIMO Multiple-Input and Multiple-Output

MAC Medium Access Control

RTS Request To Send

CTS Clear To Send

AP Access Point

TDMA Time Division Multiple Access

WLAN Wireless Local Area Network

WPAN Wireless Personal Area Network

PNC Personal Network Coordinator

SLS Sector Level Sweep

BRP Beam Refinement Phase

CBAP Contention Based Access Periods

DIFS DCF Inter-Frame Space

TXOP Transmit Opportunity

NAV Network Allocation Vector

FST Fast Session Transfer

SIFS Short Inter-Frame Space

EIRP Equivalent Isotropically Radiated Power

FCC Federal Communications Commission

SINR Signal-to-Interference-plus-Noise Ratio

CW Contentioned Window

QoS Quality-of-Service

DLMAC Decentralized Learning Medium Access Control

BinDLMAC Binary Search Decentralized Learning Medium Access Control

MDMAC Memory-guided Directional MAC

Chapter 1

Introduction

With the rapid growth in the evolution of wireless technologies, billions of users shifted their dependency from conventional wired devices (i.e., personal computer, fixed phone) to more convenient wireless devices such as tablet computers, and smartphones. Recent years observe a tremendous increase towards the demand for high bandwidth applications such as video streaming and news updates as well as high data exchange applications. As a result, these applications are driving an insatiable demand for wireless capacity. Further, the evolution of the fifth generation wireless systems (5G) is expected to support multi-gigabit-per-second data transmission rates. However, the scarcity of wireless resources has become a main problem for reliable and efficient communications. Thus, this calls for the development of enhanced resource allocation techniques which ensure high spectral efficiency. This thesis is dedicated to investigate such techniques to meet the ever-increasing requirement for the scarce and costly wireless resources.

One of the main challenges towards achieving efficient communication is the unpredictable variation of the wireless channel. This challenge can be addressed with adaptive resource allocation mechanisms. Indeed, the design of efficient resource allocation schemes that can counter channel variations is an ongoing effort for decades now. The main resource allocation schemes for enhancing the efficiency of the wireless channel utilization are opportunistic schedulers [4]. Opportunistic schedulers have been proven beneficial for unicast communication when users share the same channel and sequentially receive data via the corresponding unicast flow from the base station. In this case, the central scheduler of the base station will, upon knowing the channel quality of the users, transmit to the user that has the most favorable channel condition to optimize a given utility function.

Within the past decades, opportunistic scheduling has already been considered for multicasting. This is due to the capability of multicasting, which efficiently utilizes the channel resources by means of transmitting the same content to multiple receivers *simultaneously*. Although applying opportunistic scheduling may seem trivial, it may not necessarily complement multicasting and might not meet the expected performance improvement and objective of a wireless system. Unlike a unicast scheduler, which chooses a transmit rate based on a single user, a base station

with a multicast scheduler must account for the channel quality of multiple users and transmits the same frame to them with the same rate. This issue becomes more challenging in the presence of heterogenous channel conditions since both instantaneous and average channel qualities became important decision parameters. In short, multicast scheduling is inherently much more complex than unicast scheduling.

Since multicast rates are often constrained by the channel condition of the worst user, we also consider beamforming techniques. Beamforming is capable to adjust the antenna gain depending on the direction of interest. This allows improving the channel quality of the receivers with worst receive rates at the expense of worsening the channel quality of the receivers with better receive rate. The first part of this thesis addresses the complexity of designing an optimal opportunistic scheduler for multicasting in various scenarios. We design such a scheduler for both a system with and without beamforming.

Further, beamforming techniques are also used in technologies operating at frequencies beyond tens of gigahertz (i.e., mm-Wave communication) to achieve multi-gigabits-per-second data rates. Indeed, mm-Wave communication is the future for both legacy LTE and WiFi technologies as it offers a large swath of frequency spectrum. In addition, it potentially solve the limited capacity problem in meeting the ever-growing demand for wireless resources. However, the physical characteristics of wireless channels at such high frequencies demand for specialized scheduling algorithms. The conventional schedulers in LTE and WiFi for sub-5-GHz bands leverage broadcast channels to exchange synchronization and signalling information. However, mm-Wave lacks such benefit since signal at such a high frequency band experience severe attenuation and path loss. Therefore, directional transmission is required at least at one communicating end. Further, omnidirectional transmission is unreliable at such frequencies. This issue severely impairs the carrier sensing capability. Thus, the most commonly used contention-based channel access techniques from legacy WiFi communications are unsuitable for mm-Wave communication. As a result, mm-Wave communication requires channel access and scheduling schemes that operate independent of legacy carrier sensing schemes. The second part of this thesis addresses this problem and studies the underlying challenges in mm-Wave communications. Based on these insights, we then present a novel channel access control and scheduling mechanism.

1.1. Motivations

The dominant challenge in multicast scheduling roots from the mismatch of the channel qualities and the achievable data rate at each individual user within a multicast group. A Base Station (BS) transmitting at a higher rate than the maximum rate of a user causes a decoding failure. Conversely, a lower rate leads to inefficient channel utilization. Given that all users in a multicast group must be served with a single transmission in a given time slot, determining this rate is indeed challenging. The most basic wireless multicasting is broadcasting. Here the BS simultaneously transmits to all users at the rate determined by the worst user. As compared to

broadcasting, Opportunistic Multicast Scheduling (OMS) exploits channel diversity to improve overall throughput by transmitting at a rate higher than the broadcast rate. With OMS, the key decisions are twofold: (i) which users to transmit to, and (ii) at what rate should the BS transmits to the multicast group. The objective is to transmit to as many users as possible while ensuring high throughput. Since this problem is suboptimal when tackled in a disjoint manner, the design of an optimal multicast scheduler must ideally trade off between these decisions to achieve low delay, high throughput, and high fairness. Precisely, these are the objectives of solving a finite horizon problem – one of the main problem we solve in this thesis.

Regardless of the algorithm used, wireless multicasting suffers from the problem that the transmit rate is usually determined by the receiver with the worst channel quality. To overcome this problem, we can exploit composite or adaptive beamforming techniques to improve the channel quality of the worst user. A common solution for wireless multicast with beamforming is to select the pattern that maximizes the minimum rate among all receivers for a given transmit power. This technique works well in homogeneous channel conditions that is, when the long term throughput fairness is of interest. On the other hand, in heterogeneous channel conditions, users with better channel quality are served before the rest of the users. As expected, such aggressive discrimination towards users with lower channel quality leads to extremely low short term fairness. Since the transmission rate is determined by the channel quality of the selected users, the root of the problem lies in the selection of the users such that both short-term and long-term fairness are achievable. This is precisely the objective of a finite horizon problem.

Beamforming is crucial to enhance multicast throughput as it boosts the SNR at a specific direction of interest. This feature is particularly important for mm-Wave communications because radio signals at such high frequencies experience high attenuation and transmission losses. In particular, directional beamforming is used to further enhance the signal quality as well as the propagation distance. While directional transmission benefits throughput, its directional beam only covers a specific area. This leads to a problem where users fail to overhear the ongoing transmission when they are located outside of the coverage area. This problem is commonly known as deafness problem. Further, the weak omnidirectional sensing of mm-Wave impedes carrier sensing. This is particularly detrimental for the CSMA/CA access method since failure in overhearing the ongoing transmission causes collisions, which consequently impacts efficiency. This results in erratic deferral behavior and increased collisions, which lead to reduced fairness in terms of channel access and channel utilization. Therefore, it is vital to rethink the channel access protocol in order to fully exploit the abundant spectrum at extremely high frequencies.

In addition to a tailored random channel access method, mm-Wave communication also requires redesigning of its scheduling mechanism in order to circumvent the deafness problem caused by directional transmissions. This is because, without omnidirectional capability, disseminating global scheduling information using narrow beamwidth in a sequential manner (i.e., sector by sector) introduces excessive delay, potential disruption in communication, and low Quality-of-Service (QoS). Therefore, the design of quasi- or fully-distributed scheduling mechanisms for

mm-Wave that are independent of carrier sensing or sequential beamforming is of high relevance.

1.2. Contributions

This thesis proposes solutions which enhance the utilization of wireless network resources to cope with high bandwidth demand.

Firstly, in Chapter 3, we present opportunistic scheduling for erasure-coded finite horizon multicasting. The objective is to design an algorithm which ensures high throughput and low delay, as well as high fairness. To achieve these objectives, the main contributions in this chapter are as follows:

1. To obtained the optimal solution, we first formulate the finite horizon OMS problem as a *Dynamic Programming*(*Dyn-Prog*) problem. This solution optimally adapts the Modulation and Coding Scheme (MCS) to minimize the *completion time*, that is the time at which all receivers have successfully received the required amount of data.

2. The high complexity of *Dyn-Prog* (i.e., the complexity increase exponential with the number of users and channel instances) renders this approach unsuitable for many practical scenarios. However, it serves as an intuitive baseline for the development of a heuristic method. We thus propose a simple state-based heuristic that selects the MCS that maximizes the instantaneous throughput of the receiver with the lowest number of packets received so far, which we called *Max-Min*.

3. Due to the suboptimality of *Max-Min* as a result of not considering the channel states, we further design a more complex adaptive algorithm. This algorithm selects the MCS that results in the expected future system state that has the lowest expected completion time.

4. To deal with imperfect state information, we further include in the above heuristic design an estimation algorithm called *Weighted-CTe*. This algorithm estimates the distribution of the current channel state based on outdated past feedback information and predicts the evolution of receiver states.

5. Additionally, we implement an energy model for wireless multicasting to evaluate and analyze the energy efficiency of the different algorithms.

Since beamforming improves multicast rate by distributing more energy towards the weaker users in the system, Chapter 4 presents an opportunistic beamforming mechanism for finite horizon multicasting that minimizes the completion time. The main contributions are:

1. We first formulate the problem as a dynamic programming problem. This allows us to study the characteristics and tradeoffs that should be taken into account to obtain the optimal beamforming pattern. In particular, this formulation allows us to study the impact

of receiver state, instantaneous channel conditions, and average channels on the optimum receiver set to transmit to.

2. We then conduct a study of the tradeoffs in a toy scenario with two users. We use the insights that we get from this toy scenario to design a low complexity heuristic algorithm called FH-OMB that captures the main characteristics of the optimum solution.

3. However, the complexity of FH-OMB increases with the number of users. To cope with scenarios where larger receiver sets exist, the receivers are grouped according to their channel state and relative quality of the instantaneous channel. In order to further reduce the number of combinations, a group can only be scheduled if all groups that have better relative channel quality and at the same time have lower or equal receiver state are also scheduled.

4. To ensure practicability, we investigate the impact of imperfect feedback on the performance of our algorithms. To improve performance when feedback is infrequent, we also present algorithms to estimate the channel and receiver state.

In Chapter 6, we introduce a multi-band approach. We aim to solve the deafness problem in directional communications and achieve a fair and efficient contention-based access in IEEE 802.11ad mm-Wave networks. The main contributions are as follows:

1. As an initial step, we identify and analyze the impact of deafness on CSMA/CA in the 60 GHz band.

2. To address unreliable carrier sensing in the 60 GHz band, we propose a dual-band solution that couples the wireless interface on the 60 GHz band with the interface for legacy WiFi frequencies. This shifts the exchange of control messages onto a legacy IEEE 802.11 channel with lower bandwidth but wider coverage. This frees up channel time for high throughput transmission on 60 GHz.

3. Further, the proposed dual-band solution solves the deafness problem and increases MAC efficiency by exploiting omni-directional transmissions using legacy WiFi.

Lastly, in Chapter 7, we propose a self-organized scheduling approach for a multi-hop 60 GHz networks. Independent of other frequency band, this efficient scheduling algorithm copes with the deafness problem and learns the scheduling through the channel access events. The contributions are as follows:

1. We first identify the learning aspects that trigger conflict-free directional transmissions. This allows us to the maximize the airtime usage among neighbors.

2. To design a flexible scheduler that deals with network dynamics, due to traffic and channel variation, we introduce operations in an unslotted channel in which each slot comprises multiple micro-slots rather than a fixed length slot. In comparison to the slotted approach,

using an unslotted channel enables an allocation that perfectly fit the required channel time. Precisely, each user is allocated with a different channel time based on its traffic demands as well as the data rate of the link.

3. Although using an unslotted channel does cope with the network dynamics, it fails to minimize the idle time between transmission. To ensure efficient scheduling, a transmitter initiates a backward probing procedure upon a successful transmission. As a result, we achieve a scheduler with minimal inter-transmission idle time.

4. Lastly, we present a binary search mechanism to further reduce the idle transmission time between adjacent allocations.

Part I

Opportunistic Multicast Scheduling

Chapter 2

Introduction to Multicast and Finite Horizon Problem

2.1. Multicast

Multicast is a technique to distribute information from a single source to multiple destinations simultaneously. In recent years, multicasting data to mobile users (e.g., for the purpose of video streaming, video conferencing, Internet Protocol Television (IPTV), distribution of news and alerts, or application and operating system updates) has gained in importance. As an example, the most recent mobile network architecture, LTE, includes the Evolved Multimedia Broadcast Multicast Service (eMBMS) specifically for the purpose of distributing data and mobile TV content in a cellular network. Since the amount of such traffic in cellular networks is increasing rapidly and wireless resources are scarce and costly, improving the efficiency of wireless multicast is of high practical relevance.

The simplest and most common method for wireless multicasting is broadcasting. The BS transmits at some fixed low rate or the rate supported by the worst receiver to ensure that all receivers are able to receive the multicast transmission. This exploits the wireless broadcast gain, allowing a single transmission to simultaneously serve all receivers. Opportunistic Multicast Scheduling (OMS) is later introduced, and it improves over plain broadcast by exploiting multiuser diversity [49]. Having the BS transmits at a rate higher than the broadcast rate to the subset of receivers that can receive at this rate may improve overall throughput and minimize broadcast delay [39]. The intuition is that in an environment with variable channels, receivers that are not served in one slot due to bad channel conditions will be served in later slots when their conditions improve, and thus over time all receivers will eventually receive all the data. This method is commonly known as a method that maximizes the multiuser diversity gain. Hence, there is a tradeoff between multicast gain (serving all the receivers) and multiuser diversity gain (serving only the receivers which maximizes the instantaneous throughput). Specifically, whenever the BS transmits a packet, it is necessary to select a suitable transmit rate, i.e., MCS. Based on the

channel state of the receiver's link, the transmit MCS determines the amount of data transmitted per time slot as well as the packet loss probability. More robust MCSs transport less data and are more likely to be decodable.

Selecting the transmission rate and thus the subset of receivers to multicast to is a complex problem that has been the focus of a range of OMS algorithms [1]. To simplify the scheduling problem and improve performance when multicasting data, such algorithms often use erasure codes to ensure that with high probability, each packet received by a receiver is useful (unless the receiver already decoded all data) [35], i.e., the identity of the received packets is unimportant. Fixed rate Low-Density Parity Check (LDPC) [59] or rateless Luby Transform (LT) or Raptor codes [60] are examples of such erasure codes that work well in practice.

2.2. Finite Horizon Problem

In practice, multicasting data with a size on the order of only hundreds or several thousands of packets is much more common, particularly for mobile networks. For example, mobile applications, news, and operating system updates often have a size of hundreds of kilobytes to several megabytes. Also, when streaming video, it is common to apply erasure coding to blocks consisting of one or several Groups of Pictures (GOP) [17]. A GOP usually comprises a few hundred packets, depending on the video rate. Such block sizes are a suitable tradeoff between coding efficiency and playout delay (caused by the decoding delay). In the following two chapters, we focus on seeking the method to solve this tradeoff such that all users receive similar and fair quality of service. This problem is commonly known as the *finite horizon problem*.

The main objective of a *finite horizon* multicast problem is to minimize the completion time, i.e., the time needed for all receivers to receive enough data to decode the current block. The relevant optimization criterion is thus the throughput achieved over the duration of a block, rather than the long-term average throughput which is considered for the infinite horizon problem. This makes finite horizon multicast problem inherently more complex than the infinite horizon counterpart. When multicasting an infinite amount of data among a homogeneous group of receivers (i.e., with the same average channel conditions), the optimum tradeoff between multiuser diversity and multicast gain only depends on the number of receivers and their current channel conditions. In expectation, differences in the amount of data received by the different receivers will even out over time and therefore do not have to be taken into account. As a result, it is possible to exploit opportunistic gain (multiuser diversity gain) very aggressively, since lagging receivers have an infinite amount of time to catch up.

In contrast, the optimum decision in the *finite horizon* case depends on the amount of data received thus far. More accurately, it depends on the amount of data each receiver still needs to obtain in order to decode the full block of data and thus complete the reception. Intuitively, in case a receiver lags behind in the reception when many other receivers are also still far from completing the block, the lagging receiver may catch up by itself and jointly maximizing throughput for all

receivers may be the optimum decision. If, however, all other receivers are close to completion, optimizing the MCS only for the lagging receiver may be the optimum choice to minimize overall completion time, as all other receivers are likely to complete before the lagging receiver in any case. Precisely, the decision when to exploit opportunistic gain and when to favor lagging users very much depends on the state of the receivers (i.e., the amount of data received thus far and therefore how close the receivers are to finishing). In addition, the higher the number of users, the higher the multi-user diversity and therefore the potential for opportunistic gain. Exploiting opportunistic gain aggressively leads to higher average rates in the short term, but at the same time may lead to some users finishing early, thus reducing potential future opportunistic gain. In summary, having to take into consideration the receiver state in the optimization substantially changes the problem compared to the infinite horizon problem. Further, it also makes it much more challenging to find optimal solutions for the finite-horizon multicasting problem.

In the following chapters (Chapter 3 and Chapter 4), we design scheduling algorithm focusing on solving the finite horizon problem for multicast scheduling and multicast beamforming, respectively. Similar to prior work (e.g., [35]) we assume erasure coding of transmitted data, which is highly beneficial in wireless multicast scenarios and ensures that each packet received by a receiver is, with high probability, useful.

Chapter 3

Opportunistic Scheduling for Finite Horizon Multicasting

3.1. Introduction

Most of the existing OMS algorithms [35,38,39,49,50,75] consider the *infinite-horizon* multicast problem, where the sender has an infinite number of packets to send. The goal of the optimization is thus to maximize the *long-term* throughput to *all* receivers. This setting is a good approximation for the case of multicasting very large files [71]. In the infinite-horizon problem, the average channel quality seen by an individual receiver is likely to be close to the actual average of the channel distribution (law of large numbers). Therefore, it is unlikely to see large differences in the number of received packets among the receivers (over a sufficiently large amount of time) and the state of the system in terms of the number of received packets can be neglected. Nevertheless, as explained before, this is only true for the case where the receivers are homogeneous (having the similar average channel). Otherwise, it is not optimal even when it is evaluated over a long time interval.

In a nutshell, depending on the scenario, algorithms that are optimal for the infinite-horizon problem may be far from optimal for the finite-horizon case.

The main contributions of this chapter are as follows:

1. We formalize the finite horizon OMS problem and propose a *Dynamic Programming* based solution that optimally adapts the MCS to minimize the *completion time*, the time at which all receivers have successfully received the required amount of data.

2. The high complexity of *Dyn-Prog* renders this approach unsuitable for many practical scenarios. We thus propose a very simple state-based heuristic that selects the MCS that maximizes the instantaneous throughput of the receiver with the lowest number of packets received so far (called *Max-Min*).

3. We further design a more complex adaptive algorithm that selects the MCS that results

in the expected future system state that has the lowest expected completion time. We estimate completion time using a weighted Euclidean distance metric. The corresponding *Weighted-CT* algorithm measures the distance between the different possible future states and the final state where all receivers completed, with weights based on average throughput estimates of the receivers. To deal with imperfect state information, we further design an estimation-based version of the algorithm, *Weighted-CTe*. It estimates the distribution of the current channel state based on outdated past feedback information and predicts the evolution of receiver states.

4. We compare the performance of our low-complexity heuristics to the optimal *Dyn-Prog* solution as well as to existing approaches that greedily maximize the throughput for all receivers and a broadcasting scheme that always transmits to all receivers. We analyze scenarios with homogeneous and heterogeneous receiver sets under a basic multi-state channel model as well as Rayleigh fading. Under Rayleigh fading and with up to 16 users, the *Max-Min* algorithm provides a performance gain of 95% over the broadcasting scheme and a gain of 15% over the throughput maximization scheme. At a slight increase in complexity, the *Weighted-CT* heuristic performs very close to the optimal *Dyn-Prog* strategy, with performance gains of 120% and 30% over the broadcast and throughput maximization schemes, respectively. For scenarios with more than 16 users the achieved gains are even higher. In scenarios with imperfect information, *Weighted-CTe* achieves gains of up to 130% and 60% over the prior schemes in homogeneous and heterogeneous scenarios, respectively. We further implement an energy model for wireless multicasting and analyze the energy efficiency of the different algorithms.

This chapter is organized as follows. Section 3.2 reviews prior work. In Section 3.3 we discuss our system model, including the channel model and the MCS dependent packet loss model. In Section 3.4, we provide the dynamic programming-based optimal scheduling algorithm, together with a basic example to provide some intuition into how the algorithm trades off receiver throughput depending on the system state. To address the problem of state space explosion and high complexity of the dynamic programming solution, we propose two low-complexity heuristics in Section 3.5. Simulation results that compare the different algorithms in terms of completion time and energy consumption are presented in Section 3.6. Finally Section 3.7 concludes this chapter.

3.2. Related Work

The idea of OMS was pioneered by Gopala and Gamal [49] who studied the tradeoff between multiuser diversity and multicast gain. They analyzed the performance of three different scheduling mechanisms that adapt the transmit rate to the user with the best channel, the worst channel, and the median channel, respectively. In their follow-up work [50], they investigated the perfor-

mance achieved by serving a fixed fraction of users. This restriction is relaxed in [20] and [35]. In [20], Ge *et al.* initiate a transmission if the multicast threshold is satisfied. As the threshold is pre-determined, the transmission rate at each slot is fixed. This limits achievable throughput in case all the receivers have good channel conditions in a slot and higher rate could be used. Kozat *et al.* show that dynamic selection ratios that select more than 50% of the users can achieve higher throughput [35]. They also present an algorithm where the user selection ratio depends on the channels of the receivers. Both works presented in [20] and [35] exploit erasure coding for reliable transmission.

The authors of [38] propose algorithms with a static selection ratio (fixed for all transmissions) and a dynamic selection ratio (adapted to the instantaneous channels for each transmission) that maximize overall throughput. In [39], the authors extend their work of [38] from homogeneous to heterogeneous scenarios, composed of different groups of homogeneous users. A similar optimization algorithm for multicast throughput maximization is proposed in [75]. While all of these works target the infinite horizon case, in [41] the authors consider scenarios with a finite number of multicast packets. Using extreme value theory, they derive the optimal selection ratio for each transmission that minimizes completion time. In contrast to our work, their optimization algorithm does not consider the state of the receivers in terms of the number of received packets. Wang *et al.* [73] consider both channel and receiver state for the finite horizon problem using a disjoint formulation technique in which, in a slot, the MCS is chosen to serve the user or users with the highest priority, which is determined based on instantaneous throughput and remaining packets. However, the objective of this work is to improve the fairness between the users rather than optimizing for throughput or completion time.

All of the above literatures use a simple outage based channel model, where packet errors are deterministic. Receivers with channel conditions better than some threshold receive the packet and all other receivers lose the packet. In real wireless systems, packet errors are much more random and depend on noise and interference. In our model, we explicitly take the relationship between the channel conditions, the chosen MCS, and the probability of error into account. Also Ho *et al.* [26] take the probability of error into account in their formulation for maximizing the opportunistic multicasting gain. However, their schemes is a simple instantaneous throughput maximization and is thus not suitable for the finite horizon problem. In summary, all of the above algorithms – except those of [41] and [73] – focus on the infinite-horizon scenario and are sub-optimal for the finite-horizon case we consider in this chapter.

The problem of minimizing the overall delay for all users to receive a certain number of packets is studied in [74] through a dynamic programming approach. This work does not consider erasure coding over a larger block of data, but repeatedly multicasts a single packet until each receiver has obtained it. The BS then multicasts the next packet in the same manner, and so on. The goal of the optimization algorithm is therefore to minimize the number of transmissions required to multicast a single packet to all receivers, and the state of the system is the number of receivers that did not yet receive the packet. The algorithm adapts its decision to the changes in

the set of users that still need to receive the packet and maximizes the throughput for those users. The approach is mainly suitable for a single homogeneous group of users, since its complexity increases exponentially with the number of user groups in heterogeneous scenarios. The method of multicasting a single packet repeatedly is also much less efficient than multicasting blocks of erasure coded packets as is done above.

The most basic scheme against which we compare our proposed algorithms is the *Broadcast* algorithm (called Least Channel Gain (LCG) user rate in [1]), where the transmission rate is limited by the receiver that currently has the worst channel. This scheme ensures successful transmission to all receivers, but may sacrifice a lot of throughput when channels are highly variable. We further compare against a scheme called *Greedy* that optimizes the selection ratio at each transmission opportunistically based on the current channel states of all receivers so as to maximize total throughput. This mechanism has a performance that is indicative of the different selection ratio based throughput maximization algorithms above.

Opportunistic multicast algorithm highly depends on the availability of the instantaneous channel information from the users. The impact of limited feedback is examined in [26] and [29]. In [26], the MCS decision is made based on the average SNR. This assumption is too conservative since a realistic channel is also characterized by the path loss, Rayleigh fading and shadowing. Huang *et al.* [29] determine the transmission rate for the opportunistic multicast scheduling in [41] based on the recent channel conditions instead of the average SNR. Recent channels, however, do not always accurately reflect the instantaneous channel and it highly depends on the feedback interval, which is not explicitly defined.

In summary, our main contribution over prior work is the optimization of completion time for finite horizon multicasting. We further model the packet error rate rather than just considering outage, and evaluate the proposed algorithms under multipath Rayleigh fading. Lastly, we also propose an estimation based algorithm that accounts for imperfect receiver and channel state information.

3.3. System Model

We model the system as a time-slotted broadcast system with a single BS and N mobile users within the coverage area of the BS. Each user must receive a block of data of B bits, called the block size. We assume that due to erasure coding, each packet transmitted by the BS and received by a user is useful if the user has received less than B bits. In case multiple blocks of data are to be transmitted in succession, the BS will start transmitting the next data block only after all the receivers received the current block.

A time slot is of fixed duration. Thus, the BS broadcasts a fixed number of symbols per slot, which – depending on the MCS – corresponds to a variable number of bits. We assume that the BS can select one of the M MCSs, indexed by $m = 1, \ldots, M$. The number of bits per slot that can be transmitted using MCS m is denoted by R_m.

Perfect Channel State Information (CSI) and knowledge of the number of bits a user has successfully received is assumed to be available at the base station prior to the transmission in each time slot. The users see independent channel instances $h_i[k]$ at each time slot k. The discrete-time channel model for the received signal $s_i^{\text{rx}}[k]$ at user i is given by:

$$s_i^{\text{rx}}[k] = h_i[k]s^{\text{tx}}[k] + n_i[k], \tag{3.1}$$

where $s^{\text{tx}}[k]$ is the signal broadcast from the BS at time slot k, and $n_i[k]$ is additive white Gaussian noise term with power spectral density N_0.

For the analysis, and in particular for dynamic programming solution, we assume a discrete set of C possible channel realizations $\mathcal{H}_i$ for user i. However, the heuristics we develop also work with continuous channels. The probability of user i seeing channel coefficient $h_i \in \mathcal{H}_i$ in slot k is $\alpha_i(h_i)$, i.e., $\alpha_i(h_i) = \text{P}(h_i[k] = h_i)$. The corresponding SNR is denoted by γ_{h_i}. The vector of channels of all users, also referred to as the channel combination, is denoted by $\mathbf{h} = \{h_i, i = 1 \ldots, N\}$. We denote by $\mathcal{H}$ the set of all possible channel combinations, and by $\alpha(\mathbf{h}) = \prod_{i=1}^{N} \alpha_i(h_i)$, the probability of a channel combination $\mathbf{h} \in \mathcal{H}$. Note that the total number of channel combinations is C^N.

In contrast to prior work, we do not assume deterministic channel outage but use the PER for a given channel quality and MCS from [42]. For a channel instance $h_i \in \mathcal{H}_i$, the PER for user i under MCS m is represented by $p_i^m(h_i)$, and the corresponding Packet Success Rate (PSR) is $q_i^m(h_i) = 1 - p_i^m(h_i)$.

In this chapter, we use the following terms: *(1)* A *strategy* g specifies the MCS $g(\mathbf{h})$ for each channel combination $\mathbf{h} \in \mathcal{H}$. Hence, the total number of strategies is $S = M^{C^N}$. We denote by $\mathcal{G}$, the set of all possible strategies. *(2)* The *state* consists of the vector of the number of bits received by each user i, denoted by $\mathbf{x} = \{x_i, i = 1, \ldots, N\}$. The state space $\mathcal{X}$ consists of all states where the number of bits received by all users is positive and less than or equal to B.[1] The *initial state* where none of the users have any information is $\mathbf{x}^0$ and the end state where all the users have received B bits is denoted by $\mathbf{x}^B$. *(3)* A *policy* μ maps any given state $\mathbf{x} \in \mathcal{X}$ to the strategy $g_{\mathbf{x}}^{\mu}$ to be used in that state. *(4)* The *expected completion time* $D_\mu(\mathbf{x})$ is the mean time required to get from state $\mathbf{x}$ to the end state $\mathbf{x}^B$ under policy μ.

3.4. Optimization Problem

In this section, we consider the case of memoryless channels and formulate the problem as a stochastic shortest path problem [7] with cost per stage equal to 1 (the time needed per slot is fixed, $\tau = 1$) and no terminal cost. We assume that the probability of successfully receiving

[1] Note that to reduce the state space, we can use a normalized block size that is measured in units of the greatest common divisor of the MCS rates instead of in bits together with a corresponding normalized rate. As an example, for a block size of $B = 1800$kbits, $M = 2$ and MCSs with rates of 6Mbps and 9Mbps, the normalized block size is $B = 600$ and the normalized rate is $R_1 = 2$ and $R_2 = 3$, respectively.

a packet is non-zero for every combination of MCS and channel condition, though it might be extremely low for some combinations.

3.4.1. Dynamic programming solution (*Dyn-Prog*)

Let $\mathcal{E} = \left\{ \mathbf{e} \,\middle|\, |\mathbf{e}| = N, e_i \in \{0,1\} \right\}$ be the set of all vectors of size N whose components take values 0 or 1. The transition probability from state $\mathbf{x} \in \mathcal{X}$ to state $\mathbf{y} \in \mathcal{X}$ when MCS m is used under channel combination $\mathbf{h}$ is given by:

$$\rho_{\mathbf{h}}^{m}(\mathbf{x},\mathbf{y}) = \sum_{\substack{\min(\mathbf{x}+R_m\mathbf{e},B)=\mathbf{y} \\ \mathbf{e}\in\mathcal{E}}} \left(\prod_{i=1}^{N} p_i^m(h_i)^{e_i} q_i^m(h_i)^{1-e_i} \right), \tag{3.2}$$

where the above minimization is defined element-wise. Note that in case every user experiences an erasure, the state remains unchanged.

The state space is finite, and there clearly exists a finite integer K such that there is a positive probability of terminating after K steps irrespective of the policy. Thus, the optimal policy μ^* satisfies Bellman's equations for every state $\mathbf{x}$:

$$D_{\mu^*}(\mathbf{x}) = \min_{g\in\mathcal{G}} \left(\tau + \sum_{\mathbf{h}\in\mathcal{H}} \alpha(\mathbf{h}) \sum_{\mathbf{y}\in\mathcal{X}} \rho_{\mathbf{h}}^{g(\mathbf{h})}(\mathbf{x},\mathbf{y}) D_{\mu^*}(\mathbf{y}) \right), \tag{3.3}$$

and the optimal strategy in state $\mathbf{x}$ is given by

$$g^{\mu^*}(\mathbf{x}) = \operatorname*{argmin}_{g\in\mathcal{G}} \left(\sum_{\mathbf{h}\in\mathcal{H}} \alpha(\mathbf{h}) \sum_{\mathbf{y}\in\mathcal{X}} \rho_{\mathbf{h}}^{g(\mathbf{h})}(\mathbf{x},\mathbf{y}) D_{\mu^*}(\mathbf{y}) \right). \tag{3.4}$$

Since the state space is finite, there are several options to solve for the optimal policy as well as the minimum expected completion time. We choose a simple value iteration approach. Starting from the end state $\mathbf{x}^B$, we use Bellmann's equation Eq. 3.3 to determine the completion times of the states that only depend on the end state (for which the completion time is known to be 0). We then proceed in the same manner to determine the expected completion times of states that only depend on states for which the completion time is already known, until the completion times for all states are computed. This process also yields the optimum policies from Eq. 4.5.

3.4.2. A simple two user example

Consider a scenario with two users ($N = 2$), with identically distributed channels. Let $\mathcal{H}_1 = \mathcal{H}_2 = \{L, H\}$, and $\mathcal{H} = \{HH, HL, LH, LL\}$, where L and H denote channels with low and high channel quality, respectively. The base station can choose one of three MCSs in each slot. The probability of packet error when MCS m is used is denoted by $p^m(L)$ and $p^m(H)$ for both

users under the low and high channel, respectively. A strategy is defined by specifying the MCS to be used for each vector channel in $\mathcal{H}$.

Here, Bellman's equation at state $\{x_1, x_2\}$ is:

$$\begin{aligned}
&D_{\mu^*}(\{x_1, x_2\}) = \\
&\tau + \min_{g\in\mathcal{G}} \sum_{\mathbf{h}\in\mathcal{H}} \alpha(\mathbf{h})\Big(p^{g(\mathbf{h})}(h_1)p^{g(\mathbf{h})}(h_2)D_{\mu^*}(\{x_1, x_2\}) \\
&\quad + p^{g(\mathbf{h})}(h_1)q^{g(\mathbf{h})}(h_2)D_{\mu^*}(\{x_1, \min(x_2 + R_{g(\mathbf{h})}, B)\}) \\
&\quad + q^{g(\mathbf{h})}(h_1)p^{g(\mathbf{h})}(h_2)D_{\mu^*}(\{\min(x_1 + R_{g(\mathbf{h})}, B), x_2)\}) \\
&\quad + q^{g(\mathbf{h})}(h_1)q^{g(\mathbf{h})}(h_2)D_{\mu^*}(\{\min(x_1 + R_{g(\mathbf{h})}, B), \min(x_2 + R_{g(\mathbf{h})}, B)\})\Big).
\end{aligned}$$

We evaluate the optimal policy in a scenario where the H and L channels for the users are $\mathcal{H}_1 = \mathcal{H}_2 = \{5\text{dB}, 28\text{dB}\}$. The probabilities of L and H are $\alpha_1(L) = \alpha_2(L) = 0.75$ and $\alpha_1(H) = \alpha_2(H) = 0.25$. We choose such a highly variable channel, since it makes it easier to demonstrate the tradeoffs that the algorithm makes in the different regions of the state space. The probability of each channel combination in $\mathcal{H}$ can be easily obtained by multiplying the respective channel probabilities. For simplicity, we use $M = 3$ MCSs with normalized rates of $R_1 = 1$, $R_2 = 4$ and $R_3 = 9$. The PER for each MCS and channel instance is listed in Table 3.1.

Table 3.1: PER for different MCS and SNR value pairs

$p_i^m(h_i)$	$m = 1$	$m = 2$	$m = 3$
$\gamma_H = 28dB$	0	0	0.08
$\gamma_L = 5dB$	0.23	0.97	1

In Fig. 3.1, a drift vector (arrow) reflects the optimal policy at that state. It shows the expected future state, given the optimum MCSs chosen for the different channel combinations. Hence, the length of a vector indicates the throughput obtained by the corresponding policy. (Note that for better readability, we only plot policy vectors for a subset of states and increase their lengths.) We set the normalized block size $B = 200$. At the initial state $\mathbf{x}^0 = \{0, 0\}$ the optimal policy is $g(\mathbf{h}) = \{1, 3, 3, 3\}$, i.e., MCS $m = 1$ is used for channel combination LL and MCS $m = 3$ is used for channel combinations LH, HL, and HH. This particular policy is a greedy policy which gives the maximum throughput to both users. This policy is also used in almost all states up to $\mathbf{x} = \{140, 140\}$. Closer to the boundaries of the state space, the policy changes from greedy to increasingly favoring the user that is lagging behind. This accounts for the fact that the leading user is likely to finish before the trailing user, even if MCS decisions are optimized for the trailing user. It aims to prevent the loss of multiuser diversity caused by one user finishing early. For $\{140 < x_1 \leq 160, x_2 < 140\}$ and $\{x_1 < 140, 140 < x_2 \leq 160\}$, the predominant policies are $g(\mathbf{h}) = \{1, 3, 2, 3\}$ and $g(\mathbf{h}) = \{1, 2, 3, 3\}$, respectively, where a more conservative MCS is chosen when the trailing user has a low channel quality. This policy sacrifices throughput to prevent the trailing user from falling further behind. Even closer to the boundaries, the policies

are $g(\mathbf{h}) = \{1, 3, 1, 3\}$ and $g(\mathbf{h}) = \{1, 1, 3, 3\}$, further trading off overall throughput for a higher packet reception probability for the trailing user. When both users received a similar number of bits and are close to the end state $\mathbf{x}^B$, the algorithm also chooses a more conservative MCS indicated by a shorter arrow length to avoid overshooting (i.e., unnecessarily delivering more than B bits to both users).

While solving the stochastic shortest path problem minimizes average completion time and provides the optimal policy, the size of the state space and the computational complexity increase exponentially with N, the number of users. Therefore, the above approach is not practical for actual implementation in a network.

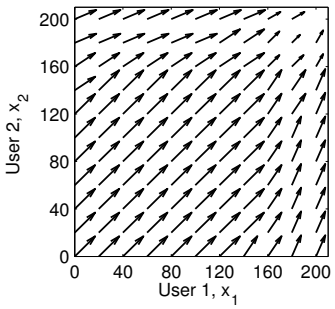


Figure 3.1: Policy given by the (*Dyn-Prog*) algorithm

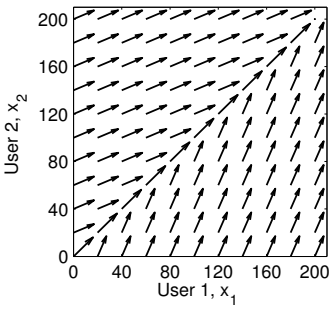


Figure 3.2: Policy given by the *Max-Min* algorithm

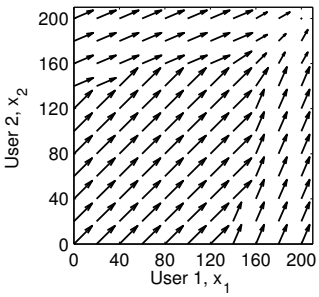


Figure 3.3: Policy given by the *Weighted-CT* algorithm

3.5. State-Aware Heuristics

Due to the high complexity of the *Dyn-Prog* algorithm introduced in the previous section, we propose two low-complexity heuristics that mimic its behavior.

3.5.1. Maximize minimum throughput (*Max-Min*) for the trailing user

This heuristic is based only on the current state and the current channel conditions. At each slot, the user with the least number of received bits is identified as the worst user, who is most likely to require the highest number of slots to receive all data. Note that this is indeed the case when the users are homogeneous and have an identical channel distributions. In the case of heterogeneous users, the users that have channel conditions that are worse (on average) are more likely to be the trailing users and thus most likely to finish last. In each slot, the algorithm uses the MCS that maximizes the throughput for the trailing user. If both users have the same number of bits, the algorithm greedily maximizes sum throughput for both. Note that since the algorithm ignores by how much the worst user is trailing, it may react too conservatively in case the difference between users is small.

Fig. 3.2 depicts the average drift resulting from such a policy in a scenario with the same parameter setting as explained in Section 3.4.2 for homogeneous users. There are two pre-

dominant strategies that are used for all states off the diagonal. As *Max-Min* sacrifices overall throughput in favor of the trailing user as soon as a user falls behind. The resulting strategies are $g(\mathbf{h}) = \{1, 3, 1, 3\}$ and $g(\mathbf{h}) = \{1, 1, 3, 3\}$. On the diagonal, *Max-Min*'s sum throughput maximization leads to the same strategy as in the *Dyn-Prog* solution, except for the last states before finishing. Since in contrast to *Dyn-Prog*, *Max-Min* does not explicitly take expected completion time into account, it does not switch to more conservative symmetric strategies of $g(\mathbf{h}) = \{1, 2, 2, 2\}$ and $g(\mathbf{h}) = \{1, 1, 1, 1\}$, respectively. The latter would deliver (with a lower packet loss probability) just the required number of bits to finish, compared to using the highest MCS $m = 3$, which may deliver more bits than necessary or result in packet loss.

Overall, we note that compared to the optimal *Dyn-Prog* algorithm, *Max-Min* is more conservative and ensures that the progression of state is with high probability along the diagonal where both users have the same number of bits.

3.5.2. Weighted completion time (*Weighted-CT*)

In many cases, favoring the trailing user is overly conservative. In particular, when the number of pending bits is large for all users and the relative lag is small, the probability that the currently trailing user actually finishes last may be small. We now present a heuristic that more closely models the decisions taken by the *Dyn-Prog* algorithm to achieve a better tradeoff between instantaneous sum throughput and balancing the number of pending bits (i.e., the state) for the different users.

At a given slot k, with state $\mathbf{x}$ and channel $\mathbf{h}$, we evaluate the average drift and determine the expected next state, $\mathbf{y}^m$, when using MCS m as:

$$y_i^m = x_i + q_i^m(h_i[k])R_m,\ i = 1, \ldots, N \tag{3.5}$$

Note that the expected state may be real valued. We then estimate the additional time required, on average, for all users to receive B bits given that they are in state $\mathbf{y}^m$. Since computing the actual estimated remaining completion time is computationally intensive as discussed in the previous section, we use a weighted Euclidean distance metric instead. The distance is taken as the bits required for completion, divided by a weight that reflects the average rate at which a user progresses through the state space. The estimated completion time $\tau_{\mathbf{y}^m}$ is thus:

$$\tau_{\mathbf{y}^m} = \sqrt{\sum_{i=1}^{N} \left(\frac{B - y_i^m}{w_i} \right)^2} \tag{3.6}$$

and the chosen optimum MCS is $m^* = \arg\min_m \tau_{\mathbf{y}^m}$.

We choose weight w_i that is proportional to the average throughput achieved by the user under

a hypothetical policy that chooses the MCS that maximizes the rate of each channel:

$$w_i = \sum_{h_i \in \mathcal{H}_i} \max_m \alpha_i(h_i) q_i^m(h_i) R_m \,. \tag{3.7}$$

As the actual choice of MCSs in the algorithm in turn depends on the policy (and thus on the states as well as the channels of the other users), it is hard to determine the true average rate. At the same time, it is not necessary to estimate this rate very accurately; it is only necessary to obtain approximately the right relative differences in user's completion time estimates that lead to the correct choice of MCS m^*. The hypothetical policy to determine the weights above is a very simple but effective method to capture these approximate relative throughput differences among the users.

In practical scenarios, the channel distribution of individual users may not be known in advance. Further, the channel statistics of a mobile user may change over time. In such settings, we use an exponentially weighted moving average to track the user's weight. Given an instantaneous channel instance $h_i[k]$ at slot k, the estimated weight of user i, $\hat{w}_i[k]$, is given by:

$$\hat{w}_i[k] = (1-\beta)\hat{w}_i[k-1] + \beta \sum_{m=1}^{M} q_i^m(h_i[k]) R_m \,, \tag{3.8}$$

where β is a sufficiently small constant.

Fig. 3.3 shows the drifts for the *Weighted-CT* algorithm. Our choice of weights indeed captures well the relative desirability of the different states. While the set of strategies is not as rich as with the *Dyn-Prog* approach – in particular at the transition between the greedy throughput maximization strategy and the more conservative border strategies – the strategies in most of the state space are almost the same. Most importantly, this is true for states around the diagonal *which are much more likely to occur in reality than states far off the diagonal* where the number of bits for the two users differs a lot.

3.5.3. Weighted-CT with rate estimation (*Weighted-CTe*)

In the previous sections, we assumed that perfect channel and state information is available at the BS. In practice, however, feedback from receivers is delayed and reporting channel and receiver states information incurs overhead and can thus only be done periodically. To deal with such imperfect and outdated information, we design an estimation-based version of the algorithm, *Weighted-CTe*. It estimates the probability distribution of the current channel state based on the outdated past feedback and predicts the evolution of receiver states.

Assume that σ is the delay between the actual channel measurement at user i and the use of that information at the sender. The sender can now estimate the probability of the *current* channel

$h_i[k]$ having the channel state h_i, given the outdated channel information $h_i[k-\sigma]$, as

$$\hat{\alpha}_i^k(h_i) = \mathrm{P}\left(h_i[k] = h_i \in \mathcal{H}_i \mid h_i[k-\sigma]\right) . \tag{3.9}$$

The estimated reception probability of user i when MCS m is used is:

$$\hat{q}_i^m(h_i[k-\sigma]) = \sum_{h_i \in \mathcal{H}_i} \hat{\alpha}_i^k(h_i) q_i^m(h_i), \tag{3.10}$$

With this, Eq. 3.5, Eq. 3.7 and Eq. 3.8 can be rewritten, replacing the actual reception probability for a known channel $q_i^m(h_i[k])$ with the estimated reception probability $\hat{q}_i^m(h_i[k-\sigma])$.

3.6. Results

In this section, we evaluate the performance of our proposed algorithms in homogeneous and heterogeneous user scenarios and compare them to the existing *Broadcast* and *Greedy* schemes discussed in Section 3.2. For simple scenarios ($N = 2$ and $C = 2$), we also compare our results to the optimal *Dyn-Prog* solution. We start with simple scenarios to provide an intuition for the algorithms that helps to better understand the more complex scenarios. We then study the impact of block size B and the number of users N on the performance of the algorithms under multipath Rayleigh fading channels with the ITU Pedestrian B path loss model [54]. The Doppler frequency is 10 Hz and the coherence time is $t_c = 40$ ms. Given that multicast traffic will be sent concurrently with other unicast data traffic, only some of the slots of an LTE frame can be used for multicast. In the simulations with Rayleigh fading, we use two out of the ten slots (or sub-frames) of a frame for multicast. Each slot has a duration of 1 ms and thus the Transmission Time Interval (TTI) is equivalent to 5 ms. Finally, we analyze the performance in a more practical scenario with limited and imperfect feedback from the users.

The algorithms are evaluated with two performance metrics: the system completion time D (in all scenarios) and the energy consumption (in the multipath Rayleigh fading scenarios). The completion time D corresponds to the time required for all of the users in the system to receive the whole block of data. It is measured in slots of 1 ms. The energy consumption is measured in Joule.

In all of the simulations, we consider three modulation schemes: Quadrature Phase Shift Keying (QPSK), 16-Quadrature Amplitude Keying (QAM), and 64-QAM, with channel coding and data rates that are corresponding to Channel Quality Indicator (CQI)=3 to CQI=15 in [42]. The MCS determines the number of transmitted bits and the PER for the instantaneous channel quality in a slot. The block size B used throughout this section is 6400kbits unless otherwise specified. This block size roughly corresponds to the size of a GOP for a video with Digital Versatile Disc (DVD) quality [16].

3.6.1. Completion time comparison to the optimal *Dyn-Prog* solution in simple scenarios

In this section, we first analyze the performance of the different algorithms in a simple $N = 2$ user scenario with $C = 2$ channel instances and $M = 13$ MCSs (CQI $= 3$ to CQI $= 15$ in [42]). In such a simple scenario, it is possible to obtain the optimum *Dyn-Prog* solution. The system model that we use here is the one described in Section 3.4.2. We analyze both homogeneous and heterogeneous user scenarios. In these scenarios, we set the stationary channel probabilities for the H and L channel to $\alpha(H) = 0.25$ and $\alpha(L) = 0.75$, respectively.

3.6.1.1. Homogeneous network

In the homogeneous scenario, both users have the same channel statistics but independent channel instances. We present the results of increasing the channel variability δ, where δ is the SNR difference between the H and L channel of each user. Here, the lowest δ is 0.7dB ($\gamma_H = 9.0$dB and $\gamma_L = 8.3$dB) and the highest δ is 20.0dB ($\gamma_H = 21.0$dB and $\gamma_L = 1.0$dB). The H and L channel pair of a user is picked such that the average throughput the user would obtain in a single-user scenario does not change. As δ increases, the SNR of the H channel increases and the SNR of the L channel decreases.

Fig. 3.4 shows how channel variation impacts the performance of the algorithms in a homogeneous scenario. In this scenario, both *Greedy* and *Weighted-CT* perform close to the optimal *Dyn-Prog* solution. As the users have the same channel distribution, exploiting opportunistic gain and maximizing the instantaneous throughput as *Greedy* does is a good strategy. The low complexity heuristic, *Weighted-CT*, weighs the homogeneous users equally according to Eq. 3.7 and transmits with the MCS that leads to the expected future state with the lowest completion time using Eq. 3.5 and Eq. 3.6. *Weighted-CT* trades off the instantaneous opportunistic throughput and the homogeneity of the receiver states to minimize the completion time in a manner similar to *Dyn-Prog*. It thus performs slightly closer to *Dyn-Prog* than *Greedy*. In contrast to *Greedy*, *Weighted-CT* exploits opportunistic multicast less aggressively in case this leads to one of the users trailing too far behind. However, since the users are homogeneous, this does not happen often and therefore the performance differences are small.

Broadcast performs worse than the other schemes because its transmission rate is limited by the lowest instantaneous channel. The impact is more pronounced when δ is larger since the SNR of the L channel is lower. Generally, *Max-Min* performs worse than the other schemes but *Broadcast*. In case the trailing user has a better instantaneous channel than the other user, *Max-Min* transmits at a higher rate than *Broadcast*, which is beneficial in a homogeneous scenario since exploiting opportunistic gain is a good strategy. However, when the trailing user has a worse instantaneous channel, *Max-Min* may send at the broadcast rate, and thus performs worse than *Dyn-Prog*, *Weighted-CT*, and *Greedy*. For $\delta = 3.7$dB, transmitting at the broadcast rate is the right decision. Here, all schemes transmit at the broadcast rate except for *Max-Min*, which

thus performs worse.

Note that the completion time of *Greedy*, *Weighted-CT* and *Dyn-Prog* is not constant – it first increases, then slightly decreases – although the hypothetical single user throughput would be equal for the different δ, as described in the setup above. This change is due to the fact that the algorithms select the transmit rate depending on the channel instances of both users and thus the distribution of transmit rates is different from the single-user case.[2]

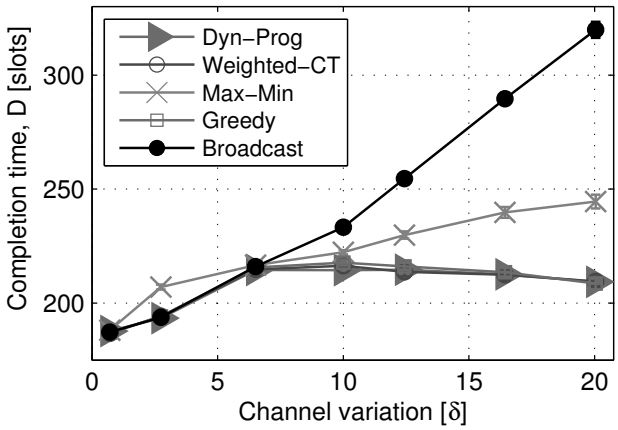


Figure 3.4: Homogeneous network with increasing channel variability δ for $N = 2$ and $B = 6400$kbits.

3.6.1.2. Heterogeneous network

In a heterogeneous channel scenario, the *good* user (g) has a better *average* channel than the *bad* user (b). In Fig. 3.5, we evaluate the completion time when increasing the average SNR of the *good* user $\bar{\gamma}^g$ from 2.8dB to 22.8dB and fixing the average SNR of the *bad* user $\bar{\gamma}^b$ at 2.8dB. Higher average SNR results in a higher rate, thus as an overall trend the completion time decreases. Here, the difference between each the H and L channels of both users is always $\delta = 12.4$dB.

On the left extreme of Fig. 3.5 (at $\bar{\gamma}^g = 2.8$dB), the users are homogeneous and the relative performance of the algorithms is as discussed in Section 3.6.1.1. *Weighted-CT* performs close to the optimal *Dyn-Prog* for all $\bar{\gamma}^g$. This confirms that also for heterogeneous scenarios, the computation of y_i^m and $\tau_{\mathbf{y}}^m$ based on the average rate estimate w_i leads to MCS decisions almost identical to those of *Dyn-Prog*. When $\bar{\gamma}^g$ is low (users are homogeneous), it maximizes aggregate user throughput, whereas when $\bar{\gamma}^g$ is high (users are heterogeneous), it is more conservative towards the *bad* user.

[2]For example, in the single-user case, $\mathcal{H} = \{H, L\}$ and the probability to transmit at the rate that corresponds to H is $\alpha(H) = 0.25$ and $\alpha(L) = 0.75$ for L. In the two-user case, $\mathcal{H} = \{HH, HL, LH, LL\}$. Here, maximizing the opportunistic gain is optimal and the probability to transmit at the rate for H is $\alpha(H) = \alpha(HH)+\alpha(HL)+\alpha(LH) = 0.44$ and at the one for L is $\alpha(L) = \alpha(LL) = 0.56$.

When the difference between the *good* and the *bad* user is sufficiently large (at $\bar{\gamma}^g = 8.8$dB), *Greedy*'s completion time increases drastically and for higher $\bar{\gamma}^g$ it performs worse than the other algorithms. For these high channel differences, serving the good user alone provides a higher sum throughput than serving both users. Therefore *Greedy* serves users sequentially – first the *good* user at a high rate until the user finishes and only then the *bad* user. In contrast, *Broadcast* and *Max-Min* schemes which always favor the trailing user outperform the *Greedy* scheme. *Broadcast* is the best strategy when it is optimal to only serve the user with the worst channel (for $\bar{\gamma}^g \geq 14.8$dB). At $\bar{\gamma}^g = 12.8$dB, *Max-Min* performs slightly worse than *Broadcast* because with some small probability the *good* user may still trail, causing the algorithm to transmit at a too high rate resulting in a very low PSR at the *bad* user. *Max-Min* experiences an increase in D for $\bar{\gamma}^g = 4.8$dB for the same reason.

To further analyze the algorithms, we show the instantaneous sum throughput at each time slot, averaged over 200 simulation runs. Instantaneous sum throughput is the total throughput of the users remaining in the system in a given slot. Fig. 3.6 shows a homogeneous scenario, corresponding to $\bar{\gamma}^g = 2.8$dB in Fig. 3.5. Fig. 3.7 shows a heterogeneous scenario, corresponding to $\bar{\gamma}^g = 12.8$dB in Fig. 3.5. Note that the user throughput is zero for a user that has already received all the data.

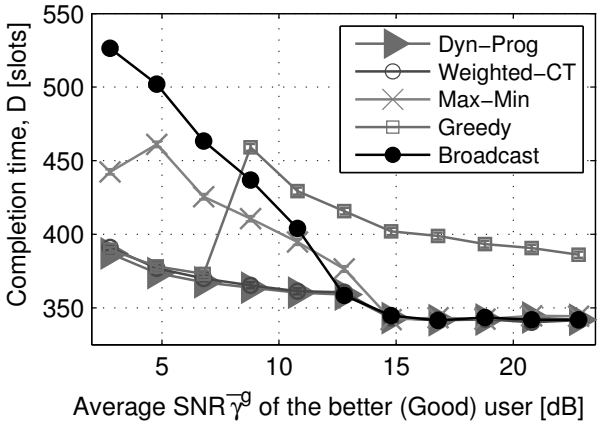


Figure 3.5: Heterogeneous network with increasing heterogeneity for $N = 2$ and $B = 6400$kbits.

It is optimal to exploit opportunistic gain and serve the better user when the users are homogeneous. Therefore, Fig. 3.6 shows that *Broadcast* achieves a lower sum throughput than the other schemes because its rate is limited by the worst instantaneous channel. Consequently, the users finish later than the other schemes (taking between 500 to 600 slots). *Max-Min* transmits at a higher rate when the trailing user has a better channel and thus achieves higher sum throughput than *Broadcast*. As discussed before, the performance of *Weighted-CT* and *Greedy* is close to the *Dyn-Prog* scheme, which is also evident in the sum throughput curves.

Fig. 3.7 shows the sum throughput of the schemes when users are heterogeneous. At time slots ≤ 100, *Greedy* has the highest sum throughput because it first transmits at a high rate to

the *good* user only, to maximize opportunistic gain. The *good* user can be served at around three times the rate of the *bad* user. Consequently the sum throughput drops significantly in time slot $\approx$ 100 slots when the *good* user finishes, leaving only the *bad* user in the system. Clearly, in this scenario serving primarily the *bad* user is the better strategy, since also the *good* user will receive those data.

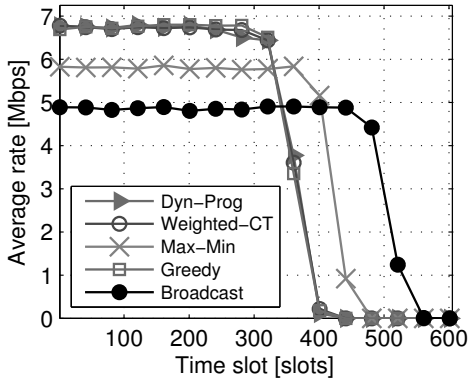


Figure 3.6: Instantaneous sum throughput for a homogeneous network.

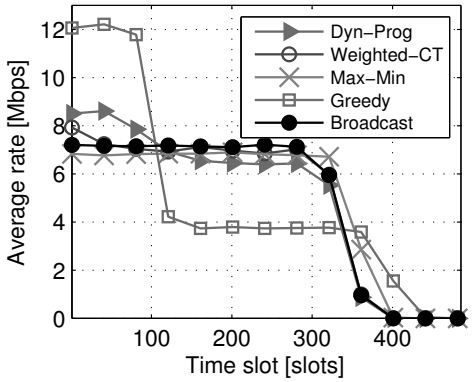


Figure 3.7: Instantaneous sum throughput for a heterogeneous network.

In contrast to *Greedy*, with *Dyn-Prog*, *Broadcast*, and *Weighted-CT* the users finish relatively close together, since the algorithms serve both users simultaneously. However, *Dyn-Prog* has a higher sum throughput than *Max-Min*, *Broadcast*, and *Weighted-CT* for time slots $\leq$ 100 slots. The optimum strategy is to opportunistically favor the *good* user *as long as both users have a small difference in terms of the amount of received data.* This effect is depicted in Fig. 3.8 where the state space visited by the *Dyn-Prog* algorithm is shown. The darker the square, the more often the corresponding states are visited. The *good* user receives more data than the *bad* user at the beginning of the transmission (i.e., when the *bad* user has 2000kbits, the *good* user usually has $\approx$ 3000kbits, but may have as few as 2000kbits). For time slots $\geq$ 100 slots in Fig. 3.7, *Dyn-Prog* transmits at the rate of the trailing *bad* user to ensure that the *bad* user catches up. Since a scheme that transmits at the rate of the *bad* user sometimes serves only the *bad* user, *Dyn-Prog* yields a lower sum throughput than *Broadcast*, *Max-Min*, and *Weighted-CT* for time slots $\geq$ 100 slots. Fig. 3.8 reflects the characteristic of *Dyn-Prog* serving the *bad* user for $x_{bad} \geq 1500$kbits and $x_{good} \geq 2500$kbits where the progress along the x-axis (*bad* user) is larger than that for the *good* user.

Weighted-CT has a lower sum throughput than *Dyn-Prog* (see Fig. 3.7) for time slots $\leq$ 100 slots because the computation of the expected completion time is sub-optimal – it does not take into account all of the (exponential number of) strategies that the optimal *Dyn-Prog* algorithm explores. From Fig. 3.9, we see that the sub-optimality of *Weighted-CT* causes it to be more conservative at the beginning of the transmission, and in turn achieves a slightly higher rate

later on since the states of the *good* and *bad* user are closer together. Nonetheless, *Weighted-CT* performs very close to *Dyn-Prog* compared to the other algorithms for all $\bar{\gamma}^g$ since *Dyn-Prog*'s slightly more aggressive initial behavior only provides a marginal reduction in completion time.

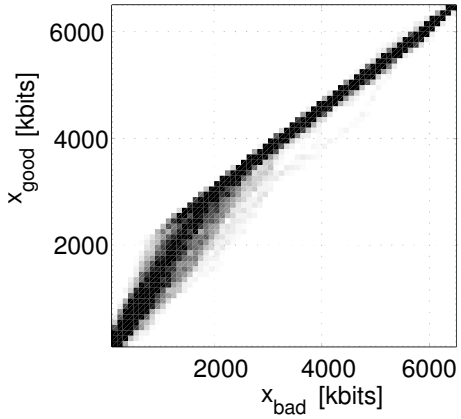


Figure 3.8: State space visits for *Dyn-Prog* at $\bar{\gamma}^g = 12.8$dB.

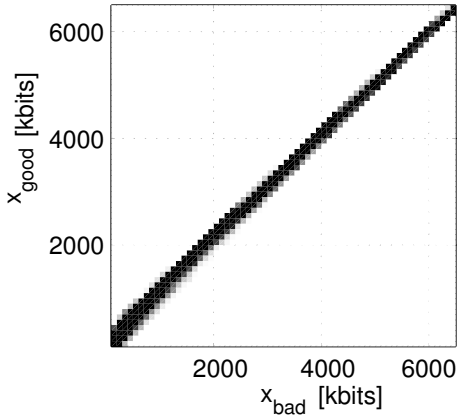


Figure 3.9: State space visits for *Weighted-CT* at $\bar{\gamma}^g = 12.8$dB.

3.6.2. Completion time comparison in multipath Rayleigh fading networks

In the following, we use the flat multipath Rayleigh fading model with path loss (ITU-R Pedestrian B [54]) and evaluate the performance in both homogeneous and heterogeneous scenarios for different numbers of users N (see Section 3.6.2.1) and block sizes B (see Section 3.6.2.2). From this section onwards, due to the complexity of *Dyn-Prog* mentioned in Section 3.5, we only evaluate the performance of *Weighted-CT*, *Greedy*, *Max-Min* and *Broadcast*.

In a homogeneous scenario, we place the users equidistant from the BS while in the heterogeneous scenario, the users are randomly distributed within the cell coverage area of radius 250m. The multipath Rayleigh channel distributions of the users are independent and identically distributed (i.i.d.). The transmit power is set such that the edge user is still able to receive data with some probability of success at the lowest MCS. Note that since we use the pedestrian channel model, the rate of change in channel SNR is rather low.

3.6.2.1. Impact of increasing the number of users N

Here, we observe the impact of increasing N exponentially from 2 to 64 and fixing B at 6400kbits. Fig. 3.10 and Fig. 3.11 show this impact for a homogeneous and heterogeneous scenario, respectively. The completion time increases with N because a higher N increases the probability that there is a user with a low SNR channel. To make it easier to compare the relative performance of the schemes, we also include a graph that shows the relative increase of the completion time for each scheme compared to the best scheme, for each scenario.

Homogeneous scenario. The channel variation (i.e., the difference between the best and the worst channel instance) of the Rayleigh fading channel is high and thus we can observe similar relative performance between the schemes in Fig. 3.10 and Fig. 3.4 (a homogeneous 2 user scenario in Section 3.6.1.1) for $\delta \geq 10$dB. As mentioned, the optimal scheme for homogeneous scenario is the one that exploits opportunism and transmits at a higher rate to achieve maximum throughput at each time instant. Therefore *Weighted-CT* and *Greedy* achieve a lower D compared to the other schemes. As N increases, which user is the trailing user changes more often and the trailing user may be a user with a better channel. Since it is better to opportunistically serve users with better channel in a homogeneous scenario, *Max-Min* performs better than *Broadcast*. *Broadcast* performs worst and its completion time increases with N because the probability that there exist a user with a very low SNR in a given slot is higher for higher N.

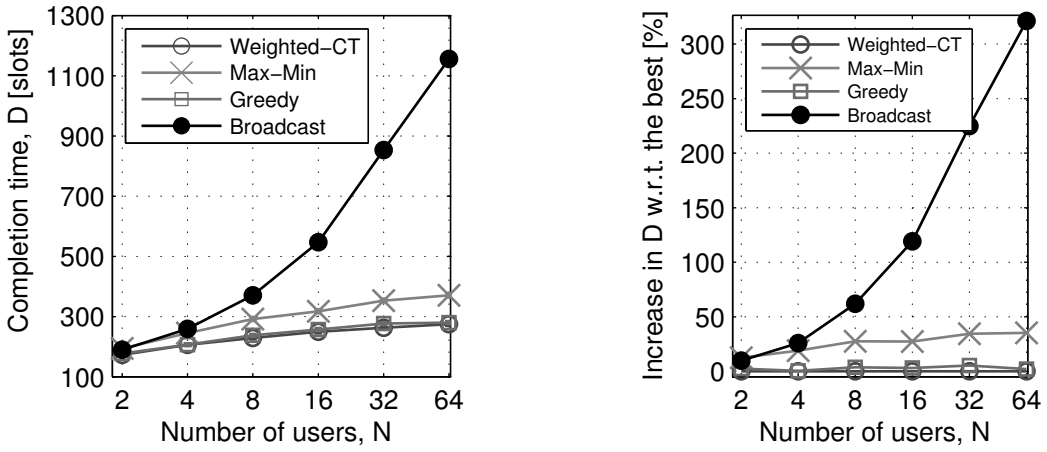


Figure 3.10: Impact of increasing N for homogeneous multipath Rayleigh fading scenario for $B = 6400$kbits.

Heterogeneous scenario. In Fig. 3.11, the heterogeneity (i.e., the difference of the average channels) between a user and its nearest neighbor is higher when the number of users in the system is small. Therefore, we observe that the relative performance of the schemes in a Rayleigh fading channel (see Fig. 3.11) for small N is similar to the performance of the heterogeneous 2 user scenario (see Fig. 3.5 in Section 3.6.1.2) for higher $\bar{\gamma}^g$. According to the explanation in Section 3.6.1.2, *Broadcast* performs close to optimal and *Greedy* that optimizes for opportunistic gain performs worst when there is one clearly worst user. As N increases, the user's density increases and thus the users that are close to each other have a very small difference in terms of the average channels. This resembles a homogeneous scenario. In such a scenario, transmitting at the broadcast rate and conservatively serving all users results in higher D than using opportunistic gain among the users that are near each other. Therefore *Greedy* performs better than *Broadcast* for higher N. For high N, the multicast rate is highly affected by the users located at the edge of a cell (i.e., edge users). Since the trailing user is normally an edge user, *Max-Min* may by chance serve the correct users. Therefore, it outperforms *Greedy* for serving the important users and

Broadcast since it is less conservative. As before, *Weighted-CT* outperforms all other schemes since it exploits the users' weight and optimizes the rates at which the edge users are served.

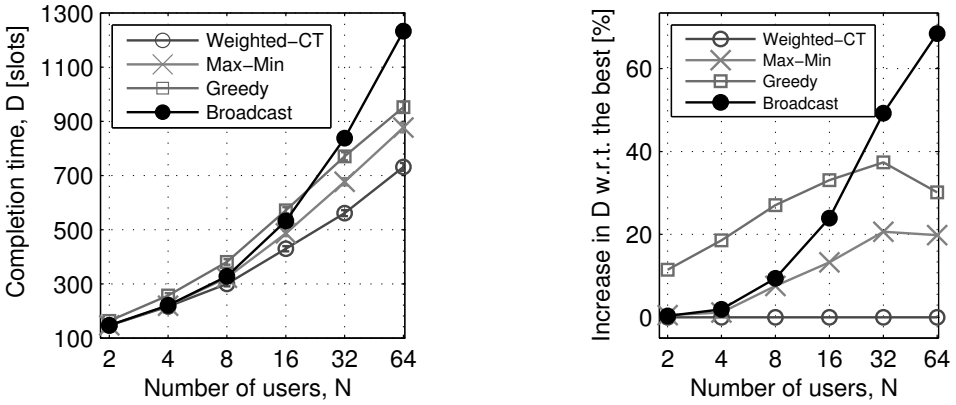


Figure 3.11: Impact of increasing N for heterogeneous multipath Rayleigh fading scenario for $B = 6400$kbits.

3.6.2.2. Impact of increasing the block size B

Here, we fix N to 16 and examine the impact of increasing B exponentially from 50kbits to 25600kbits. For ease of the comparison, the results are presented in terms of normalized completion time, $D_n = D/B$. Fig. 3.12 and Fig. 3.13 illustrate the impact of increasing B on D_n. When B is small, a scheme requires very few slots to receive a block and thus the number of the channel samples is small. Due to this, the average SNR of the channel samples and the average SNR of the channel distribution may differ significantly. Note that D is determined by the worst users in the system. If the average SNR of the channel samples of some of the users is lower than the average SNR of the channel distribution, this causes a higher D. This impact reduces for larger B. According to the law of large numbers, for larger B, the average SNR obtained from the larger number of channel samples is closer to that of the distribution. The larger D allows the system to stay in steady state (where all users are still far from finishing) for a longer period of time. As a consequence, D_n becomes flatter (i.e., the increase in D is approximately linear with B).

Homogeneous scenario. When B is small (e.g., $B = 50$kbits), the users with a better instantaneous channel are more likely to finish very early (within very few slots) and the users with a worse instantaneous channel remain in the system. In such a scenario, an algorithm performs best if it serves all the users at the rate of those with the worse channel instances. Therefore *Broadcast* performs well. *Weighted-CT* performs best since it is more conservative towards the users with a worse instantaneous channel. In contrast, *Greedy* performs worst because it at first serves users with a better instantaneous channel and only serves the users with worse instantaneous channel later. *Max-Min* selects a user among the trailing users randomly when more than one user has the lowest amount of received data. When this randomly selected trailing user is not the user with the

worst instantaneous channel, it yields a higher instantaneous throughput but this causes longer D in the future. For larger B, the performance difference is as explained in Section 3.6.2.1 for the homogeneous scenario when $N = 16$.

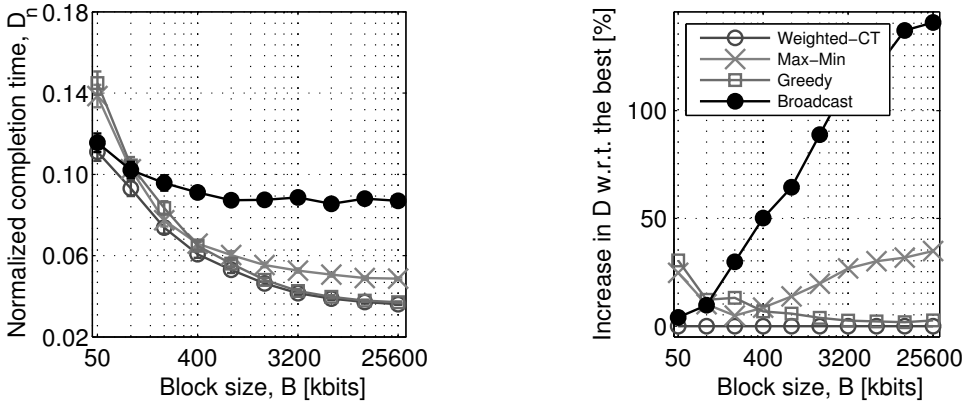


Figure 3.12: Impact of increasing B for homogeneous multipath Rayleigh fading scenario for $N = 16$.

Heterogeneous scenario. In a heterogeneous scenario, the reason for the performance difference between *Greedy*, *Max-Min*, and *Broadcast* in Fig. 3.13 for small B is similar to that explained for Fig. 3.12. *Greedy* performs worst regardless of B since it always favor the better users (refer to Section 3.6.1.2 for a detailed explanation). Since serving the worse users is important in this scenario, *Max-Min* always performs better than *Greedy* because the trailing user is one of the worse users. *Weighted-CT* performs slightly worse than *Broadcast* for $B = 50$kbits because for a very short completion time, there is a discrepancy between the estimated rate (computed using Eq. 3.8) of the channel samples and that of the channel distribution. This incorrect rate estimation causes *Weighted-CT* to make the wrong decisions (i.e., choosing a wrong MCS). For larger B, higher number of channel samples allows *Weighted-CT* to estimate the users' rate and completion time more accurately and thus it performs better than the other schemes.

3.6.3. Evaluation of the energy consumption

This section presents the energy consumption of the experimented schemes in the homogeneous and heterogenous multipath Rayleigh fading scenarios with perfect feedback.

At slot k, a user is either in the *on* state or in the *idle* state. A user is in the *on* state if it is scheduled for reception and *idle* state otherwise. The parameters for computing the energy consumption are listed in Table 3.2[3].

As expected, the energy consumption is proportional to the completion time. This can be seen in the homogeneous scenario where a scheme with the lowest completion time (*Weighted-CT*) in Fig. 3.10 (see Section 3.6.2.1) has the lowest energy consumption in Fig. 3.14.

[3] The energy consumption in LTE system is presented in [28] and [5]. We ignore the energy consumed by the wake up operation since it is less than 5% of the energy consume in an *idle* state [28].

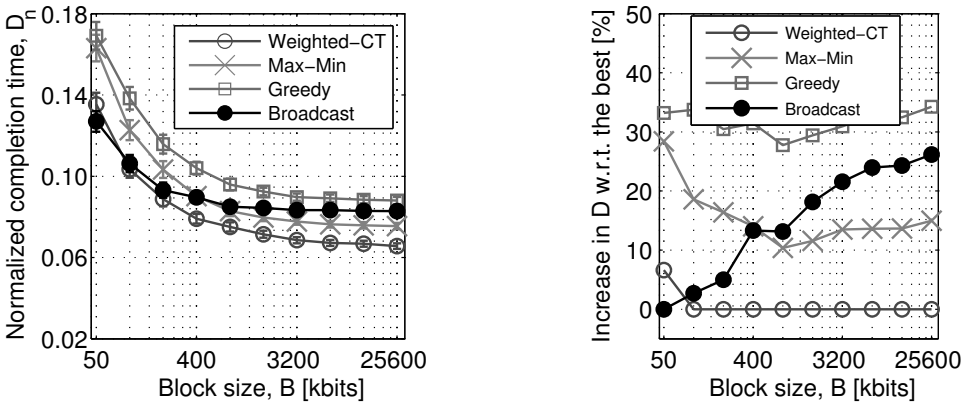


Figure 3.13: Impact of increasing B for heterogeneous multipath Rayleigh fading scenario for $N = 16$.

Table 3.2: LTE power model parameters

Symbol	Description	Value
ξ_{on}	*on*'s state base power	1210.7 ± 85.6 (mW)
ξ_{idle}	*idle*'s state base power	594.3 ± 8.7 (mW)
θ_{dl}	Power per Mbps	51.97 (mW/Mbps)

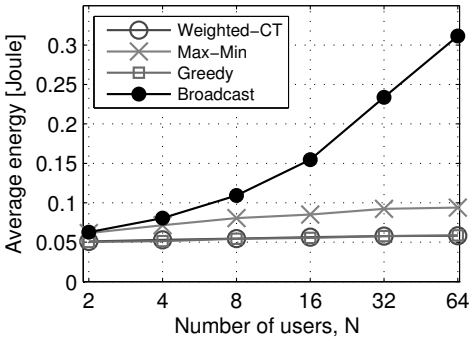


Figure 3.14: Average energy consumed in homogeneous multipath Rayleigh fading scenario for $B = 6400$kbits.

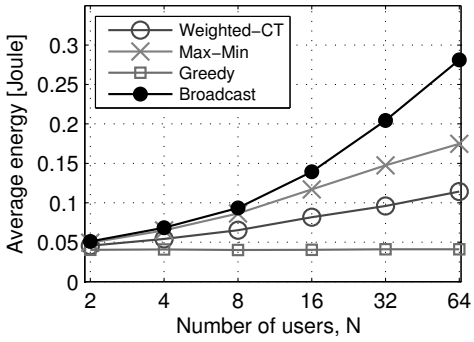


Figure 3.15: Average energy consumed in heterogeneous multipath Rayleigh fading scenario for $B = 6400$kbits.

Although *Weighted-CT* has lower completion time than *Greedy* in Fig. 3.11, its average energy consumption is higher (see Fig. 3.15). The energy consumption is not only affected by the completion time but it is also affected by the number of slots a user stays in the *on* state. Table 3.2 shows that a user in the *on* state consumes more than twice the amount of base energy than it does in the *idle* state. In a heterogeneous scenario, *Weighted-CT* consumes higher average energy than *Greedy* because it tries to serve more users and transmits at a lower rate than *Greedy*. Transmitting at lower rate causes the better users to stay in the *on* state for more slots. In contrast, *Greedy* first transmits at a high rate and reduces the number of slots the better users stay in the *on* state.

As a result, its average energy consumption is lower. Increasing N does not decrease the rates at which users are served hence the average energy consumption of *Greedy* remains the same as N increases. In contrast, for *Weighted-CT*, a higher N causes more users to stay in the system for a longer time and thus its average energy consumption increases with N in Fig. 3.15.

In summary, the average energy consumption depends on two main factors: the completion time and the percentage of the slots the users stay in the system (especially in the *on* state).

3.6.4. Impact of imperfect and limited state information

The schemes presented in this chapter need either the channel state or the receiver state information or both to select an MCS. In wireless networks, frequent feedback of the state information substantially increases system overhead and thus reduces system throughput. In this section, we examine the impact of imperfect feedback information on all schemes. We also include *Weighted-CTe* that estimates the probability distribution of the current channel state based on the outdated past feedback (see Section 3.5.3). First, we analyze the impact of increasing the feedback interval (i.e., reducing the frequency of feedback) of the state information. We then study the effect of limiting the number of users that periodically feed back their state information to the BS.

3.6.4.1. Impact of the feedback interval

As mentioned above, transmitting feedback at each slot is costly. Here, we model a feedback system where users only send feedback every λ slots. We then analyze the impact of the feedback interval λ on the completion time D. The feedback slot (the slot at which a user reports) is asynchronous, but λ is the same for all users.

We choose $\lambda \in \{5\text{ms}, 10\text{ms}, 20\text{ms}, 40\text{ms}, 80\text{ms}, 160\text{ms}\}$ (these are the common periodic feedback intervals in an LTE system [57]), such that we can observe the impact of λ when it is less than or greater than $t_c = 40ms$. For $\lambda \leq t_c$, the current channel state and the one reported in the last feedback (i.e., the last available channel state) are correlated, otherwise, they are uncorrelated. When feedback is unavailable, the BS updates the receiver state assuming that the amount of data received in the current slot is equivalent to that reported in the last available feedback packet. When feedback is available, the BS updates the receiver state to the current receiver state.

Fig. 3.16 and Fig. 3.17 depict the impact of different feedback intervals λ on D in the homogeneous and the heterogeneous scenarios, respectively. D increases with λ because the correlation between the current channel state and the last available channel state decreases and thus the uncertainty about the channel state increases. As mentioned, the receiver state is updated regardless of the availability of feedback information. The difference between the current and the estimated receiver state is usually small and thus has a minimal impact on the chosen MCS. Hence, delayed channel state has a much higher impact than the receiver state on the completion time.

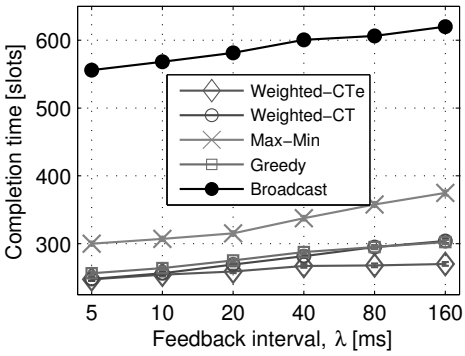


Figure 3.16: Impact of increasing λ for a homogeneous multipath Rayleigh fading scenario, $B = 6400$kbits, $N = 16$.

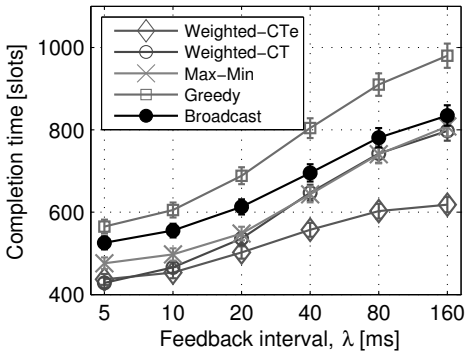


Figure 3.17: Impact of increasing λ for a heterogeneous multipath Rayleigh fading scenario, $B = 6400$kbits, $N = 16$.

Fig. 3.16 and Fig. 3.17 show that D continuously increases with λ. For $\lambda > t_c$, the last available channel state is only useful up to time t_c. From time t_c until the next feedback is received, the current and the last available channel state are uncorrelated. Therefore, increasing λ beyond t_c increases the fraction of time where the current and the last available channel state are uncorrelated. When comparing the impact of λ on D in the homogeneous (see Fig. 3.16) and the heterogeneous (see Fig. 3.17) scenarios, the impact is greater in the latter. In a homogeneous scenario (with a sufficiently large N), the difference between the chosen MCS based on the last available channel state and the one that would be chosen based on the current channel state is usually small. In contrast, in a heterogeneous scenario, an outdated channel state may cause an algorithm to select a very different MCS compared to the MCS that would be selected based on the current channel state since the MCS largely depends on the channel state of a small subset of users. Choosing either a too low or too high MCS increases D. As a result, the impact of delayed state information is higher in the heterogeneous scenario than in the homogeneous scenario. For $\lambda > t_c$, although *Weighted-CT* may make the wrong MCS choice due to outdated last feedback, it performs as good as or better than the other schemes except for *Weighted-CTe*.

Weighted-CTe outperforms the other schemes in all scenarios, especially for higher λ. The MCS that is chosen for an outdated channel is the one that is optimal for the distribution of the estimated channel (since little is know about the channel). As a result, *Weighted-CTe* that picks the MCS using channel estimation significantly outperforms all the other schemes especially for high λ. With channel estimation, the completion time of *Weighted-CTe* improves over *Weighted-CT* by 17.5% and 30% in homogeneous and heterogeneous scenarios, respectively. As mentioned, the difference of the selected MCS between the current channel state and the last available channel state is larger in the heterogeneous scenario therefore the impact of channel estimation is more evident in this scenario.

We also investigate the performance of *Broadcast*, *Max-Min*, and *Greedy* when channel estimation similar to *Weighted-CTe* is applied. Due to space limitation, we exclude the graphs.

With channel estimation, *Max-Min* and *Greedy* achieve improvements of up to 29% and 21%, respectively in terms of D compared to *Max-Min* and *Greedy* without channel estimation. In contrast, *Broadcast* with channel estimation performs worse than all the other schemes, including the *Broadcast* scheme without channel estimation. This is because the estimated channel of *Broadcast* is the worst channel in the channel distribution.

3.6.4.2. Impact of interval feedback from fewer users

In this sub-section, the feedback load is further decreased by reducing the number of reporting users (N_{rep}). We examine the minimum N_{rep} for each scheme that achieves the same D as that when all users give feedback. We first elaborate the method on the selection of the reporting users. We then present the minimum N_{rep} for $N = \{2, 4, 8, 16, 32\}$ for a feedback interval $\lambda = 20$ms.

Regardless of N, *Broadcast* decides its transmit rate based on the worst instantaneous channel. *Max-Min* decides based on the channel of the trailing user. Therefore, the minimum N_{rep} for *Broadcast* and *Max-Min* is one. Since *Weighted-CT* and *Weighted-CTe* favor the users with lower receiver state (these users require longer time to receive B), these users are selected as the reporting users. *Greedy* maximizes throughput and it highly depends on the complete instantaneous channel statistic to achieve an opportunistic gain, thus it performs better for higher N_{rep}. For *Greedy*, the reporting users are selected randomly.

In Fig. 3.18 and Fig. 3.19, we show the minimum N_{rep} required by each scheme for different N. Note that we do not show the completion time here. The completion time is similar to that depicted in Fig. 3.10 (for a homogeneous scenario) and Fig. 3.11 (for a heterogeneous scenario) since for $\lambda = 20$ms the current channel state and the last available channel state are still correlated. In Fig. 3.10, *Greedy* performs close to *Weighted-CT* but the required N_{rep} is more than twice as large as that of *Weighted-CT* (see Fig. 3.18), and consequently has higher overhead than *Weighted-CT*.

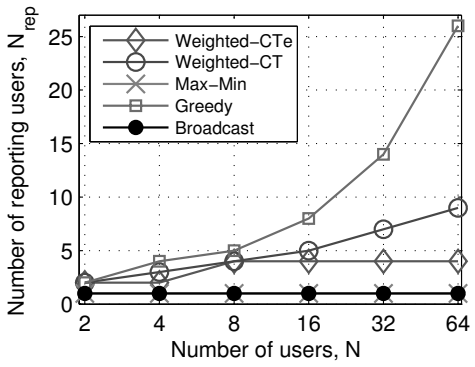


Figure 3.18: N_{rep} in a homogeneous multi-path Rayleigh fading scenario, $\lambda = 20$ms.

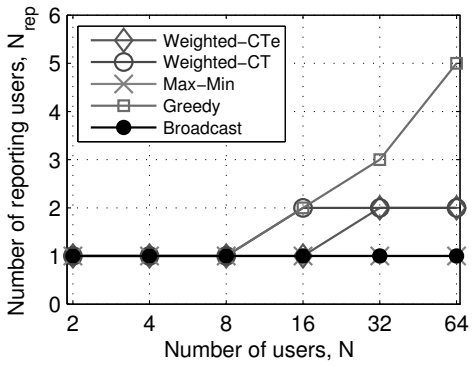


Figure 3.19: N_{rep} in a heterogeneous multi-path Rayleigh fading scenario, $\lambda = 20$ms.

As depicted in Fig. 3.18 and Fig. 3.19, N_{rep} increases with N for *Weighted-CTe*, *Weighted-CT*, and *Greedy* because more users are needed to reflect the channel statistics of the important users

for each scheme. For *Greedy* to realize the opportunistic gain, the feedback from one user is as important as any other users in the system and thus a higher N_{rep} is required.

In Fig. 3.18, it is interesting to note that the minimum N_{rep} of *Weighted-CTe* is lower than that of *Weighted-CT* in the homogeneous scenario. As explained in the previous section (Section 3.6.4.1), a high λ reduces the number of MCS choices when channel estimation is used. Therefore, only a small N_{rep} is needed for *Weighted-CTe*. On the other hand, *Weighted-CT* has a wider range of MCS choices thus including more users results in smaller D. As discussed in Section 3.6.4.1, in the heterogeneous scenario, it is important to select the edge users since the completion time is determined by those users. The number of the edge users increases with N and thus N_{rep} is larger for larger N. *Weighted-CT* and *Weighted-CTe* have the same N_{rep} in Fig. 3.19 except for $N = 16$. This small difference is caused by the number of MCS choices as explained above.

To recap, the minimum number of reporting users N_{rep} depends on the network scenario. Since users are usually randomly distributed in actual mobile networks, they are heterogeneous. As shown in our result (see Fig. 3.19), the BS (for any algorithm it uses) can achieve its best performance with low overhead because it only needs feedback from a small fraction of users in the system to achieve the minimum completion time for each scheme.

3.7. Conclusions

In this chapter, we investigated the finite horizon opportunistic multicast scheduling problem, where a wireless base station transmits a fixed amount of erasure coded data to a set of receivers. We designed an algorithm based on dynamic programming that to the best of our knowledge, is the first to explicitly take into account the system state in terms of the received amount of data at each receiver for the selection of the optimum modulation and coding scheme. In addition to the well known tradeoff between broadcast gain and multi-user diversity gain that is inherent to opportunistic multicast scheduling, the finite horizon nature of our problem introduces an interesting further tradeoff, namely that of equalizing the completion times of the users versus the total system throughput. This tradeoff is state dependent. Intuitively, throughput maximization is a reasonable strategy as long as users are far from finishing, whereas the closer users are to finishing, the more important it becomes to allow lagging users to catch up rather than optimizing throughput for all. Based on these insights, we designed two simple and practical heuristics that perform close to the more complex dynamic programming solution and that outperform existing approaches that do not consider the receiver state.

We performed an extensive range of simulations for homogeneous as well as heterogeneous user scenarios and show that our heuristics outperform the existing *Greedy* and *Broadcast* schemes by as much as 30% and 120%, respectively, in Rayleigh fading scenarios. Since reducing the system overhead improves the system throughput. We also present results on the impact of increasing the feedback interval and reducing the number of users that give feedback on the pro-

posed schemes. In particular, we extend our heuristic to estimate the current channel state based on outdated last feedback. This extension further improves performance by 17.5% and 30% in homogeneous and heterogeneous scenarios, respectively.

Chapter 4

Opportunistic Beamforming for Finite Horizon Multicasting

4.1. Introduction

Wireless multicast is an efficient technique to disseminate multimedia data to groups of users. Using the broadcast nature of the wireless medium allows data to be served to multiple users simultaneously, but at the same time this constrains the transmit rate to the rate that can be supported by the receiver with the worst channel conditions. If, however, there are some receivers that on average have worse channels than the rest, exploiting multiuser diversity and transmitting preferentially to the users with better channels by giving to the good receivers is detrimental to performance. Here, the overall rates are still limited by the worst receivers.

To overcome this problem, transmit beamforming can be used to adjust antenna gains to the different receivers. This allows improving the SNR of receivers with bad instantaneous channels at the expense of worsening those of receivers with better instantaneous channels. There are two main techniques for multi-user beamforming: (i) composite beamforming [67] and (ii) adaptive beamforming [3]. A composite beam is composed of multiple pre-determined single-lobe beam patterns. In contrast, adaptive beamforming calculates antenna weights directly based on the measured channels to the different receivers. While composite beamforming has lower complexity, adaptive beamforming may achieve better performance in multi-path rich environments. Similar to opportunistic multicast, the main challenge when designing multi-user beamforming mechanisms is the tradeoff between high gain beamforming to few receivers versus lower gain beamforming to a larger receiver set [56, 67, 78].

Most of the prior work solving the multicast beamforming problem aims at maximizing the rates of the receivers, which is optimal for the *infinite* horizon problem (i.e., where an infinite amount of data is to be sent). In contrast, we analyze the more realistic *finite* horizon problem where the BS sends a block of data of a certain size to all receivers.

We first model the problem and obtain the optimum solution via dynamic programming. This

allows us to study the impact of receiver state, instantaneous channel conditions, and average channels on the optimum receiver set to transmit to, and hence the optimum beamforming patterns. First, we study these tradeoffs in toy scenarios with two users. These insights allow us to design a low complexity heuristic algorithm that captures the main characteristics of the optimum solution and at the same time can run in real-time in practical wireless scenarios.

We then present a range of simulation results for larger scenarios with homogeneous and heterogeneous receiver sets and Rayleigh fading channels. While the complexity of the dynamic programming solution prevents us from solving those larger problem instances optimally, we see that our proposed heuristic provides significantly better performance than solutions based on broadcasting or greedily maximizing rates. Note that both the broadcast and greedy mechanisms use beamforming and take the instantaneous channel conditions into account. The greedy mechanism is thus optimal for the infinite horizon case (or problem instances with very large block sizes) in homogeneous scenarios as shown in [40]. The broadcast mechanism makes use of beamforming to maximize the minimum rate and hence does not suffer from receivers with bad channel conditions as much as conventional OMS. It corresponds to the solution in [78] that is optimal for scenarios with fixed channels but may be too conservative in case of variable channels. It is also optimal for scenarios with variable channels where the receiver set is highly heterogeneous and one receiver has a significantly worse average channel than the other receivers.

In practice, feedback arrives with a certain delay and the feedback frequency has to be set sufficiently low so as not to create excessive feedback overhead. We investigate the impact of imperfect feedback on the performance of the algorithms and extend our heuristic algorithm to make decisions based on partial information and estimated channel and receiver state.

Similar to prior work mentioned above, we assume erasure coding of transmitted data, which is highly beneficial in wireless multicast scenarios and ensures that each packet received by a receiver is useful (with high probability).

This chapter is structured as follows. A review of state-of-the-art for opportunistic multicast and multicast beamforming is given in Section 4.2. In Section 4.3, we model the finite horizon opportunistic multicast beamforming problem and provide an optimum solution based on dynamic programming. We design a low complexity heuristic, *FH-OMB*, in Section 4.4, and in Section 4.5 we compare its performance to the optimum solution and the greedy and broadcast schemes proposed in prior work for both perfect and imperfect feedback scenarios. Section 4.6 concludes the chapter and provides an outlook on future work.

4.2. Related Work

Opportunistic Multicasting: Opportunistic multicasting has been well studied for both the infinite horizon problem [35, 38, 39, 50, 75] as well as the finite horizon problem [41, 62, 63]. Among the first ideas to address the infinite horizon problem for homogeneous scenarios was to split the receivers into two groups according to their instantaneous channels and serve the group

with the better channel quality. As the composition of the group changes from slot to slot, all users have equal chances to be served [20, 49, 50]. This work was extended in [35] by optimizing the selection ratio, i.e., the size of the receiver set to transmit to. As a single pre-computed selection ratio is not always optimal, [39] and [41] propose a dynamic user selection mechanism that depends on the instantaneous channel at each transmission.

The authors of [41] solve the user selection problem for the finite horizon case using extreme value theory to minimize completion time. However, the user selection is only based on the instantaneous channel but not on the user state (i.e., the amount of data received by users). In wireless systems with packet loss, this is suboptimal since users may have received a different number of packets. The problem is addressed in [62, 63], where it is shown that the optimal solution for the finite horizon problem needs to take receiver state into account.

The main challenge of opportunistic multicasting is to cope efficiently with receivers with bad channel conditions. In this context, transmit beamforming can be used to balance the users' SNRs.

Multicast Beamforming: Multicast beamforming provides a trade off between multicast gain and beamforming gain. Beamforming to receivers with poor channel conditions improves the SNR at these receivers (but at the same time lowers SNR at other receivers). The basic algorithm proposed in [56] first transmits omnidirectionally to the receivers that have a high SNR and then beamforms sequentially to the remaining weak receivers. Better performance can be achieved by selecting the beamforming vector that maximizes the minimum SNR among all multicast receivers [13, 61]. In [67], receivers are partitioned into groups that are scheduled sequentially, which may outperform mechanisms that always beamform to all receivers. The work proposes two multicast beamforming mechanisms, one that splits power equally among all beams and one that allows for asymmetric power allocation. Both mechanisms use composite beamforming, where a multi-lobe beam pattern that serves multiple receivers is composed of multiple single-lobe beam patterns. In [78], the authors improve upon this work and provide an optimal solution for the equal power split and two different heuristics for the (NP-hard) asymmetric power allocation mechanism. Both [67] and [78] consider the finite horizon problem but do not take channel variations and opportunistic scheduling into account.

The same problem is addressed in [3] using adaptive beamforming rather than composite beamforming. Adaptive beamforming may provide better antenna gains than composite beamforming, in particular in multipath environments, but at the same time determining the optimum beamforming pattern is more complex.

Opportunistic Multicast Beamforming: There is very little existing work that jointly takes opportunistic multicast scheduling and multicast beamforming into account. A theoretical analysis of the optimum user selection ratio for opportunistic multicast beamforming using extreme value theory is provided in [40]. Once the user group is determined, the optimal beamforming pattern is the one that maximizes the minimum SNR among the users that are served. The algorithm is designed for independent and identically distributed users (i.e., homogeneous scenarios) for

the infinite horizon multicast problem. A similar work in [21] also focuses on the same infinite horizon problem and proposes an instantaneous throughput maximization algorithm. We show that these approaches are not suitable for the finite horizon multicast problem, especially for heterogeneous user distributions. Other relevant works presented in [11, 18, 58] improve throughput by exploiting the Multiple-Input and Multiple-Output (MIMO) capability, which are a different mechanism that is not the focus of this work.

This chapter differs from prior work in that it addresses *finite* horizon opportunistic multicast beamforming in homogeneous and heterogeneous scenarios and explicitly takes into account receiver state (i.e., the amount of data already received) to minimized delay.

4.3. System Model

We consider a wireless network with a single BS (or access point) and a set T of multicast receivers, with $|T| = N$. We assume the channels between the BS and the receivers are independent discrete memoryless channels.[1] Let $\mathcal{G}$ denote the set of all possible vector channels from the BS to the receivers. The probability that at a given time instant the channel vector $\mathbf{C}$ has channel gains $\mathbf{g} \in \mathcal{G}$ is given by $P(\mathbf{C} = \mathbf{g})$. Let C_i, g_i, and $\mathcal{G}_i$ denote the corresponding channel instance, gain, and set of possible channels for receiver i. As is common for opportunistic scheduling, we assume that the BS has perfect knowledge of the current channel instances, but for any future channel instances only the channel distribution is known.

The BS uses composite beamforming. The antenna array has K antenna patterns that are optimized to produce one strong single-lobe beam that covers a sector of approximately $\frac{360^0}{K}$ and that together cover the whole azimuth of 360^0. A composite beam is a multi-lobe beam pattern composed of several single-lobe beams that are transmitted simultaneously [67]. Each single-lobe beam k has a certain beam weight α_k. This weight corresponds to the fraction of the total transmit power allocated to that beam, and thus determines the SNR at the receivers covered by the beam. To ensure that the total radiated power remains unchanged, we have the constraint $\sum_k \alpha_k = 1$. Let k_i^* be the strongest single-lobe beam that covers receiver i and let $\gamma_{g_i}^{i,SLB}$ denote the SNR at that receiver when using that single-lobe beam when the channel gain is g_i. Then the SNR of that receiver for a multi-lobe beam is

$$\gamma_{g_i}^{i} = \alpha_{k_i^*} \gamma_{g_i}^{i,SLB} .$$

We consider a time-slotted model. In each time slot the BS transmits data to the receivers using a certain modulation and coding scheme (MCS) and beamforming pattern. For MCS $m \in M$, the number of bits transmitted in a slot is R_m and the corresponding packet reception probability for an SNR of γ is $p_m(\gamma)$. Note that we assume that receiver i will only be served when a multi-lobe beam is used with $\alpha_{k_i^*} \neq 0$.

[1] Note that our heuristic works for continuous channels and we provide simulation results for Rayleigh fading channels in Section 4.5.

4.3.1. Problem formulation

The BS has a block of data of size B (in bits) to transmit to all receivers. An erasure code is applied to the data before transmission, so that each data packet is useful for each receiver that receives it, as long as that receiver has obtained less than B bits so far.

The optimization problem is thus for the BS to select at each time slot the multi-lobe beam pattern with corresponding weights as well as the MCS that minimizes the expected completion time. Optimal choice of beam pattern and MCS depend on the current instantaneous channel, the probability distributions of the channels, and the amount of data received by the receivers so far.

When beamforming to a subset of receivers $T' \subseteq T$, the highest expected rate to those receivers is obtained by selecting beam weights α_k^* that maximize the minimum SNR at the receivers. Let $T'_k = \{i \in T' : k_i^* = k\} \subseteq T'$ be the subset of receivers served by beam k. The minimum SNR of receivers in T'_k for a single-lobe beam pattern and a given channel $\mathbf{g}$ is

$$\gamma_{\mathbf{g}}^{SLB}(T'_k) = \min_{i \in T'_k} \gamma_{g_i}^{i,SLB} \tag{4.1}$$

and, as shown in [78], the optimum weights for the multi-lobe beam pattern are thus given by

$$\alpha_k^* = \begin{cases} \left(\gamma_{\mathbf{g}}^{SLB}(T'_k) \sum_{j=1, T'_j \neq \emptyset}^{K} \frac{1}{\gamma_{\mathbf{g}}^{SLB}(T'_j)}\right)^{-1}, & \text{if } T'_k \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

This results in the same minimum SNRs for all lobes of the multi-lobe beam. Hence, all receivers in T' are served with the MCS that provides the highest expected rate

$$m^* = \arg\max_m R_m p_m(\alpha_k^* \gamma_{\mathbf{g}}^{SLB}(T'_k)) \,. \tag{4.3}$$

Thus, rather than optimizing over all possible beam weights, it is sufficient to optimize over all possible subsets of receivers.

Note that the algorithm in [78] always serves all receivers associated with a given beam, while this is no longer optimal for opportunistic multicast. Consider a scenario where all receivers are located in the same beam. This is the conventional OMS scenario for which it is well known that broadcasting to all users is not always optimal [35].

4.3.2. Dynamic programming solution for multicast beamforming

With this we can formulate the problem as a stochastic shortest path problem and solve it through dynamic programming [7]. The state is given by the amount of data received by the receivers so far $\mathbf{s} = [s_1 \ldots s_N], 0 \leq s_i \leq B$ and we denote the state space by $\mathcal{S}$.[2] As all time slots have the same duration, the cost per slot is 1.

[2]Given that there is a discrete set of rates R_m, many states cannot be reached and we remove these states from the state space to speed up the computation.

When multicasting to a subset T' of receivers with an instantaneous channel of $\mathbf{g}$, the transition probability from state $\mathbf{s}$ to state $\mathbf{s}'$ is

$$\rho_{\mathbf{g}}^{T'}(\mathbf{s},\mathbf{s}') = \sum_{\substack{\min(\mathbf{s}+R_{m^*}\mathbf{e},\mathbf{B})=\mathbf{s}' \\ \mathbf{e}\in\mathcal{E}}} \left(\prod_{i=1}^{N} p_{m^*}(\gamma_{g_i}^i)^{e_i}(1-p_{m^*}(\gamma_{g_i}^i))^{1-e_i} \right) \tag{4.4}$$

where the vector minimization above is element-wise. $\mathcal{E} = \left\{\mathbf{e} \in \{0,1\}^N\right\}$ is the set of binary vectors of size N and e_i is the ith element of $\mathbf{e}$, indicating whether receiver i received the packet or not. The MCS m^* is calculated according to Eq. 4.3. Eq. 4.4 takes into account all combinations of which receivers receive the packet and ensures that the state of receivers with $s_i = B$ does not change.

A policy $\mu_{\mathbf{s}} : \mathcal{G} \mapsto \bigcup_{T'\subseteq T} T'$ specifies the best subset of receivers to transmit to for any instantaneous channel $\mathbf{g}$ when in state $\mathbf{s}$. Let $\mathcal{M}$ bet the set of all possible mappings. Since the probability of terminating after a finite number of steps is positive, we can use Bellman's equation to find the optimal policy

$$\mu_{\mathbf{s}}^* = \underset{\mu_{\mathbf{s}}\in\mathcal{M}}{\arg\min} \left(\sum_{\mathbf{g}\in\mathcal{G}} P(\mathbf{C}=\mathbf{g}) \sum_{\mathbf{s}'\in\mathcal{S}} \rho_{\mathbf{g}}^{\mu_{\mathbf{s}}(\mathbf{g})}(\mathbf{s},\mathbf{s}') D^*(\mathbf{s}') \right). \tag{4.5}$$

The corresponding optimal expected completion time is

$$D^*(\mathbf{s}) = \min_{\mu_{\mathbf{s}}\in\mathcal{M}} \left(\sum_{\mathbf{g}\in\mathcal{G}} P(\mathbf{C}=\mathbf{g}) \sum_{\mathbf{s}'\in\mathcal{S}} \rho_{\mathbf{g}}^{\mu_{\mathbf{s}}(\mathbf{g})}(\mathbf{s},\mathbf{s}') D^*(\mathbf{s}') \right). \tag{4.6}$$

Given that the state space is finite we can solve the dynamic program through value iteration, starting from the final state $\mathbf{s}^B$. This optimization problem is hard and even a much simpler version of it with fixed channels (i.e., no opportunistic scheduling), as well as guaranteed packet delivery without errors is NP-hard [78].

The dynamic program has double exponential complexity. The state space has size B^N and for each state there are $2^{N^{|\mathcal{G}|}}$ policies that map each of the channel states in $\mathcal{G}$ to one of the 2^N possible multi-lobe patterns. Also $|\mathcal{G}|$ itself is exponential in N. Clearly, the dynamic program can only be solved for very small problem instances. For this reason, in the next section we design a lower complexity heuristic.

4.4. Heuristic Algorithm for Multicast Beamforming

Our proposed Finite-Horizon Opportunistic Multicast Beamforming (*FH-OMB*) heuristic has two main parts: 1) given the current instantaneous channel, computing the next states the system

could move to using the different multi-beam lobes that correspond to multicasting to the different subsets of receivers, and 2) estimating the expected completion times from those new states. The decision taken by the heuristic is then to beamform to the subset of receivers that results in moving to the state with the lowest expected completion time.

4.4.1. Instantaneous beamforming decision

Let the current state be $\mathbf{s}$ and the current instantaneous channel be $\mathbf{g}$. Assume the estimated completion times $D(\mathbf{s}')$ for all future states are known. When beamforming to $T' \subseteq T$ we can calculate $\gamma_{\mathbf{g}}^{SLB}(T'_k)$, α_k^*, and the resulting optimum MCS m^* using Equations (4.1)–(4.3). The expected future state $\mathbf{s}'(T')$ is given by

$$s'_i(T') = \min\left(s_i + R_{m^*} p_{m^*}(\gamma_{g_i}^i), B\right) \ \forall i \tag{4.7}$$

and the optimum subset of receivers T^* to beamform to is thus

$$T^* = \arg\min_{T' \subseteq T} D(\mathbf{s}'(T')) \,. \tag{4.8}$$

In contrast to the dynamic programming formulation we compute expected average future state rather than looking at all combinations of possible future states based on packet loss events. Note that this still requires minimization over a number of completion times that is exponential in the number of receivers, which can be done exhaustively for small receivers sets.

For larger receiver sets, we cluster receivers according to their state s_i and relative quality of the instantaneous channel. The rate receiver i would obtain with the current channel g_i for a single-lobe pattern is $R(i) = \max_m R_m p_m(\gamma_{g_i}^{i,SLB})$, and the average rate that is obtained under all possible channels is

$$\bar{R}(i) = \sum_{g_i \in \mathcal{G}_i} P(C_i = g_i) \max_m \left(R_m p_m(\gamma_{g_i}^{i,SLB})\right) \,. \tag{4.9}$$

The relative channel quality is $R(i)/\bar{R}(i)$. Let $0 = \xi_1 < \xi_2 < ... < \xi_U = B$ be a set of state thresholds and $0 = \theta_1 < \theta_2 < ... < \theta_V = \infty$ be a set of relative channel quality thresholds. We then group all receivers with

$$T_{uv} = \left\{i \in T : \xi_u \leq s_i < \xi_{u+1}, \theta_v \leq R(i)/\bar{R}(i) < \theta_{v+1}\right\}$$

where the total number of groups is UV. In Eq. 4.8, we now optimize over subsets $T' \subseteq T$ that include whole receiver groups (i.e., if one of the receivers in a group is included, the whole group must be included). We set the thresholds so that the receivers are distributed relatively evenly among the groups. In order to further reduce the number of combinations, a group can only be scheduled if all groups that have better relative channel quality and at the same time have lower

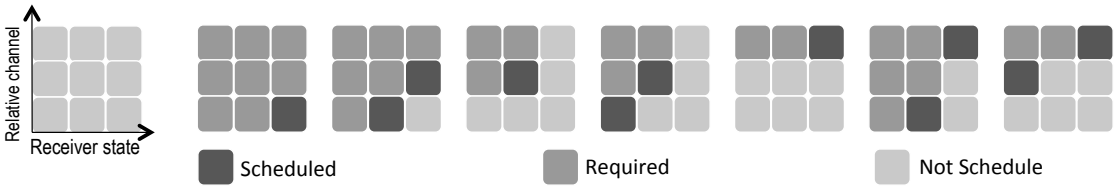


Figure 4.1: Sample for 3-by-3 user grouping. The scheduled receivers (Required) are scheduled because of the scheduled receivers (Darker block)

With this, the maximum number of combinations and thus the complexity of *FH-OMB* is

$$O\left(\frac{(V+U-1)!}{U!(V-1)!}\right). \tag{4.10}$$

The number of beamforming patterns is fixed for fixed values U and V.[3]

4.4.2. Estimating the expected completion time

The main complexity of the dynamic programming solution lies in the calculation of the expected completion time. Hence, this is what the heuristic primarily addresses. As only the instantaneous channel is known at the BS, we base the expected completion time of a future state on the average channel of the receivers. Due to the shape of the rate function, simply averaging the channel would overestimate the receive rate. Hence we first calculate the average single-lobe rate of receiver i, $\bar{R}(i)$, as given by Eq. 4.17 and then set the receiver's average SNR $\bar{\gamma}_g^{i,SLB}$ such that

$$\max_m \left(R_m p_m(\bar{\gamma}_g^{i,SLB})\right) = \bar{R}(i)\,. \tag{4.11}$$

For *fixed* SNRs and a continuous rate function, according to [78] the maximum rate when multicasting to a receiver set is obtained for a multi-lobe beam pattern that encompasses the whole receiver set. Analogous to Equations (4.1) and (4.2), for a receiver subset T' we can derive $\bar{\gamma}_{\mathbf{g}}^{SLB}(T'_k)$ as well as $\bar{\alpha}_k^*$ based on the average SNRs $\bar{\gamma}_g^{i,SLB}$ calculated above. The corresponding hypothetical average rate is given by

$$\bar{R}(T') = \bar{R}(T'_k) = \max_m R_m p_m(\bar{\alpha}_k^* \bar{\gamma}_{\mathbf{g}}^{SLB}(T'_k))\,. \tag{4.12}$$

We have $\bar{R}(T') = \bar{R}(T'_k)$ for any non-empty lobe k, since all lobes have the same minimum rate.

[3]From the simulations we find that a reasonably low value for U and V (i.e., $U = V = 4$) suffices in practice, leading to a fixed number of subsets to consider for the optimization.

With this, we can now approximate the expected completion time as follows. For a given state $\mathbf{s}$, let $T'_1 = \{i \in T : s'_i < B\}$ be the set of receivers that still require further packets and let $s'^{(1)}_{\max} = \max_{i \in T'_1} s'_i$ be the state of the receiver(s) closest to completing. When multicasting to this receiver set at rate $\bar{R}(T'_1)$ given by Eq. 4.12, one or more of the receivers would complete after a time $\tau_1 = (B - s'^{(1)}_{\max})/\bar{R}(T'_1)$. Determine the set of remaining receivers $T'_2 = \{i \in T'_1 : s'_i < s'^{(1)}_{\max}\}$ and set $s'^{(2)}_{\max} = \max_{i \in T'_2} s'_i$ to calculate τ_2, etc. In general,

$$\tau_j = (B - s'^{(j)}_{\max})/\bar{R}(T'_j)\,. \tag{4.13}$$

In other words, the estimation algorithm proceeds diagonally through the state space until hitting a boundary with $s_i = B$ for one of the dimensions, then proceeds diagonally along that boundary until hitting the next one, and so on, until reaching the final state. The algorithm terminates after at most N steps. The expected completion time is given by

$$D(\mathbf{s}') = \sum_j \tau_j\,. \tag{4.14}$$

Accounting for opportunistic gain: When determining τ_j above, we assume that receivers in T'_1 are served first, then receivers in T'_2, etc. This ignores that receiver sets will be selected based (also) on their instantaneous channels. As a consequence, $\bar{R}(T')$ is a conservative estimate of the actual rate at which this receiver group is served, since they are more likely to be served when their channel is good. We refine Eq. 4.13 to take into account opportunistic gain as follows. We assume that receivers in groups T'_1 and T'_2 are served during $\tau_1 + \tau_2$. If the channels of the receivers in $T'_1 \setminus T'_2$ are good, group T'_1 will be served, otherwise group T'_2 will be served. Hence, receivers in T'_1 see better average channels (since some of the beam weight α that was required for receivers in $T'_1 \setminus T'_2$ can now be used for other beams) whereas there is no change for receivers in T'_2. We remove the worst fraction $\tau_2/(\tau_1 + \tau_2)$ of channel combinations of the receivers in $T'_1 \setminus T'_2$ and update their average channels accordingly. We then recompute Equations (4.12) and (4.13) and obtain a new τ'_1. Similarly, the calculation of τ'_2 is based on receiver groups T'_2 and T'_3, and so on. The completion time is then calculated as $D(\mathbf{s}') = \sum_j \tau'_j$. Note that this is still a conservative estimate of the opportunistic gain.

Example and discussion:

To provide an intuition for the completion time estimation, we discuss an example for a two-receiver case in Fig. 4.2. In a two-user scenario, there are only three possible beamforming patterns serving receiver sets $\{1\}$, $\{2\}$, or both $\{1, 2\}$. For $\{1\}$ and $\{2\}$, single-lobe beam patterns with maximum array gain to the respective receiver are used, whereas for $\{1, 2\}$ the multi-lobe beam that equalizes the SNRs of the receivers is chosen. In the latter case, both receivers are served at the same rate and have the same packet loss probability. For each of the average future states $s'(\{1\})$, $s'(\{2\})$, and $s'(\{1, 2\})$ we compute the expected completion time. Consider, for

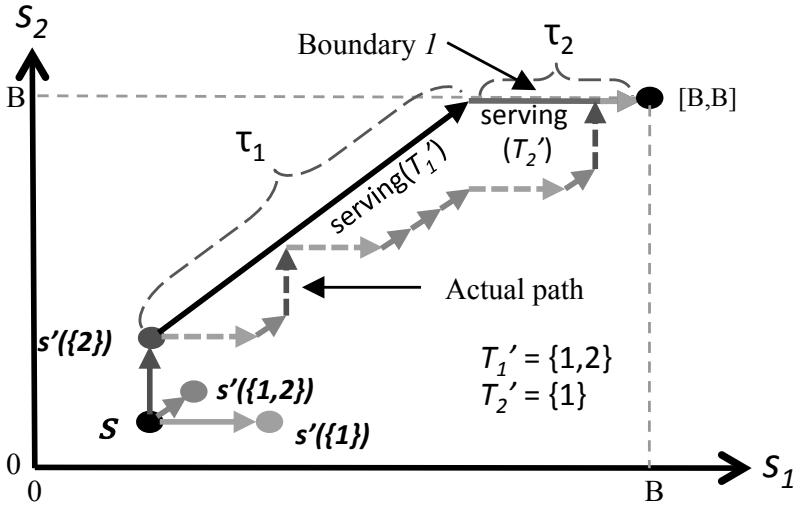


Figure 4.2: Completion time estimation with the FH-OMB heuristic.

example, $s'(\{2\})$. Since both users have not yet finished, we calculate the number of time slots τ_1 required for the first receiver to have B bits. In the example this is receiver 2. We then compute τ_2 required for the second user to complete, based on the single-lobe pattern to that user only. Using τ_2, we can recompute the first segment to obtain τ_1' that partially accounts for opportunistic gain. $\tau_2' = \tau_2$ since there is no opportunistic gain for a single receiver. The actual path that is taken through the state space (shown with a dotted line) depends on the actual instantaneous channel conditions at future states and is generally shorter than the sum of the estimated path segments.

An important observation is that determining the exact completion time is not important. What is important is to have approximately the right relative differences among completion times of nearby states (in this case $s'(\{1\})$, $s'(\{2\})$, and $s'(\{1,2\})$), such that the right *instantaneous* beamforming decisions are taken. As a consequence, it is possible to use average channels instead of all possible channel combinations, without incurring a substantial drop in performance.

4.4.3. Estimation algorithms

In the previous sections we assumed perfect information for the instantaneous channel and receiver states upon which the BS bases the decision which subset of receivers to scheduled. In practice, however, feedback always comes with a delay (at least the delay for the preparation and transmission of the feedback frame), and often even such delayed feedback may not be available for all receivers. For example, to reduce feedback overhead, feedback for a receiver may be sent only once each so-called "feedback interval", rather than in each slot. We further assume that the receivers transmit their feedback asynchronously to the BS, i.e., feedback information of different receivers is outdated by a different degree. In such cases, it is essential that the algorithm takes into account the inaccuracy of the state information, rather than directly using the feedback. To this end, we design algorithms to estimate the receiver and channel states and refine the computation of the expected completion time in Eq. 4.14 to take into account imperfect feedback.

It is important to note that the estimation of the instantaneous beam pattern (based on the

estimated channel) and that of the future beam patterns (based on the average rate) are different. To estimate the instantaneous beam pattern, the BS must first estimate the instantaneous channel based on the last known channel information and the time elapsed since this channel information was received. In contrast, to estimate the future beam pattern, it is important to estimate the average rate at which the receivers progress. However, simply averaging the estimated channels and assuming that the future rates are given by the MCS that maximizes the rate for this average estimated channel is a poor estimator of the future rate. We therefore design more sophisticated estimation algorithms that use the distribution of the estimated channels to determine the future rate and thus obtain a more precise expected completion time. The following subsections discuss the algorithms in more detail.

4.4.3.1. Receiver state estimation

Receiver state estimation is highly important to schedule the right subset of receivers, but also relatively straightforward. We estimate the expected amount of received data when up-to-date receiver state information is unavailable as follows. When a node is scheduled, it receives $R_m p_m$ packets on average, where p_m is the receive probability. R_m is determined from the beamforming pattern based on the estimated channel state, which will be explained next. In each slot, the BS assumes that a receiver receives $R_m p_m$ more packets if it is scheduled, and zero packets otherwise. Whenever the BS receives a feedback frame, it updates the receiver state with the reported correct receiver state.

4.4.3.2. Channel state estimation

Using outdated channel knowledge when the instantaneous channel information is unavailable results in inaccurate beam weights α_k^* in Eq. 4.2. Consequently, a system may wrongly boost the SNR of better receivers and vice versa. This impacts the system's performance and the problem escalates when the erroneous beam pattern causes a wrong selection of the receiver subset. In what follows, we describe an algorithm to estimate the channel in the absence of instantaneous channel information.

At a time slot t, assume that the outdated channel gain of receiver i which was reported λ slots ago (delayed by λ ms) is $g_i[t-\lambda] \in \mathcal{G}_i$. The probability that a channel C_i has a gain g_i given the outdated channel is expressed as follows:

$$\mathrm{P}\left(g_i\right) = \mathrm{P}\left(C_i = g_i, g_i \in \mathcal{G}_i \mid g_i[t-\lambda]\right). \tag{4.15}$$

Obtaining $\mathrm{P}\left(g_i\right)$ for all $g_i \in \mathcal{G}_i$ gives the probability distribution of the current channel when the feedback frame is delayed by λ. The resulting estimated channel gain is

$$\hat{g}_i = \sum_{g_i, g_i \in \mathcal{G}_i} \mathrm{P}\left(g_i\right) g_i \quad \forall \lambda = [0, \lambda_{\max}]. \tag{4.16}$$

The corresponding estimated channel instance, channel gain vector and SNR when using a single lobe beam are $\hat{C}_i$, $\hat{\mathcal{G}}_i$, and $\gamma_{\hat{g}_i}^{i,SLB}$, respectively. Note that after feedback is received, it ages in subsequent slots, until new feedback becomes available. It is therefore important to determine the channel distributions for all possible delays λ. Fig. 4.3 shows that a larger λ contributes to a wider distribution of the expected channel because a longer delay causes higher channel uncertainty.

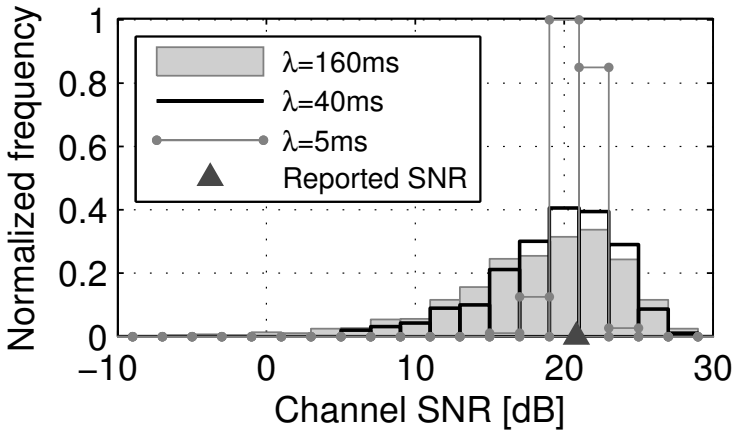


Figure 4.3: The distribution of the estimated channel SNR for different λ.

Once the $\gamma_{\hat{g}_i}^{i,SLB}$ for all receivers are obtained, we compute the optimum weights for the multi-lobe beam pattern in Eq. 4.2 by replacing $\gamma_{g_i}^{i,\text{SLB}}$ in Eq. 4.1 with $\gamma_{\hat{g}_i}^{i,SLB}$.

4.4.3.3. Completion time

Section 4.4.2 explains the computation of expected completion time taking into account opportunistic gain when perfect feedback is available. If, however, channel information is outdated, exploiting opportunistic gain becomes more difficult (i.e., the higher the channel uncertainty, the more likely it is that the perceived 'opportunity' in fact no longer exists and exploiting it would be detrimental to performance). We take this effect into account when calculating estimated completion time.

As mentioned earlier, obtaining the beam pattern to compute the expected completion time is different than that to obtain the instantaneous beam pattern. Since the distribution of the future channels is determined by the estimated channels (for different λ), the future beam pattern (which determine the progress of the receivers) is also influenced by the distribution of the received rate given by the distribution of the estimated channels. Consequently, the computation of the average future rate and the beam patterns associated with it is more complex than that in Eq. 4.12, where instantaneous feedback is available.

Computing the expected completion time requires the average rate of the receivers. To obtain the average rate, we first estimate the average channel of the receivers and then compute the beam weight that gives the average rate. However, as explained in Section 4.4.2, due to the shape of the rate function, the average estimated channel should be derived from the average rate. This

is because simply averaging the estimated channels would results in an overestimation of the average channel thus gives an inaccurate average rate.

On that account, we first compute the average estimated rate based on the distribution of the estimated channel $\hat{\mathcal{G}}_i$ from Eq. 4.16 and its corresponding MCS which maximizes the rates of each estimated channel as follows:

$$\hat{R}(i) = \frac{1}{\lambda_{\max}} \sum_{\lambda=0}^{\lambda_{\max}} \left(\sum_{\hat{g}_i \in \hat{\mathcal{G}}_i} \mathrm{P}(\hat{C}_i = \hat{g}_i) \max_m \left(R_m p_m(\gamma_{\hat{g}_i}^{i,SLB}) \right) \right) \tag{4.17}$$

To take delayed feedback into account, Eq. 4.17 includes equally distributed λ since the BS experiences an increasing delay of up to $\lambda = \lambda_{\max}$ before it receives the feedback where $\lambda = 0$ and this is repeated until the receiver has received all the intended data.

The corresponding average estimated channel $\bar{\gamma}_{\hat{g}_i}^{i,\mathrm{SLB}}$ based on $\hat{R}(i)$ that is used to determine the beam weight is then derived as follows:

$$\max_m \left(R_m p_m(\bar{\gamma}_{\hat{g}_i}^{i,SLB}) \right) = \hat{R}(i)\,. \tag{4.18}$$

The average estimated rate for a receiver subset T' is obtained by replacing $\bar{\gamma}_{\mathbf{g}}^{SLB}$ in Eq. 4.12 with $\bar{\gamma}_{\hat{\mathbf{g}}}^{SLB}$ from Eq. 4.18. Lastly, the expected completion time is computed based on the average estimated rate in Eq. 4.18 instead of $\bar{R}(T')$ in Eq. 4.12.

4.5. Results

In this section, we present simulation results to analyze the performance of the algorithms. We first investigate a simple scenario with two receivers and a two-state channel to compare the optimal dynamic programming solution (*Dyn-Prog*) and the finite horizon opportunistic multicast beamforming heuristic (*FH-OMB*) and gain insights into the optimum strategy and fundamental tradeoffs. We then investigate more realistic scenarios with multi-path Rayleigh fading channels, larger number of receivers, and larger block sizes. For these, we do not provide dynamic programming results as the run time is prohibitive due to the algorithm's complexity. The multi-path Rayleigh fading channel corresponds to the ITU Pedestrian B path loss model in [54]. For all the scenarios, we use a subset of 13MCSs given in the LTE specification for the 20MHz LTE downlink model (with modulation schemes QPSK, 16-QAM, and 64-QAM, and code rates from 0.1885 to 0.9258). The corresponding transmit rates range from 5Mbps to 95Mbps. A time slot has a duration of 1ms. The main performance metric is completion time, i.e., the number of time slots needed for all receivers to receive B kbits.

We compare the performance of *Dyn-Prog* and the *FH-OMB* heuristic with two alternative mechanisms:

1) Broadcast Algorithm: Broadcast uses a multi-lobe beam pattern that covers all receivers i with $s_i < B$, maximizes the minimum SNR across all lobes, and serves the receivers with the optimum

MCS m^* for that SNR as given in Equations (4.1)–(4.3). This scheme is presented in [78] and it is shown to be optimal for constant channels with fixed SNR.
2) Greedy Algorithm: For *Greedy*, we sort the receivers with $s_i < B$ according to their instantaneous channel quality, given by the single-lobe SNR $\gamma_{g_i}^{i,SLB}$. Let T_1 be the receiver set that includes the receiver with the best channel (that hasn't finished yet), T_2 be the set of the two receivers with the two best channels, etc. The algorithm then determines the receiver set to beamform to as

$$T^* = \arg\max_{T_j} \sum_{i \in T_j} R_{m^*} p_{m^*}(\gamma_{g_i}^{i,SLB}) .$$

The optimum receiver set is the one with the highest overall sum rate for all receivers that have not yet finished. This algorithm corresponds to the one proposed in [40] and works well for homogeneous receiver sets.

4.5.1. Simple scenario

In this section, we present the results for a simple scenario with $N = 2$ receivers and block size $B = 1000$kbits. Each receiver i has two possible instantaneous channels ($g_i = \{H_i, L_i\}$ where H and L represents the channel with a higher and lower SNR, respectively), such that $\mathcal{G} = \{H_1H_2, H_1L_2, L_1H_2, L_1L_2\}$ with $P(C_i = H_i) = P(C_i = L_i) = 0.5\,\forall i$. We analyze a homogeneous scenario and a heterogeneous scenario.

4.5.1.1. Homogeneous scenario

In this scenario receivers have the same set of channels ($H = H_1 = H_2, L = L_1 = L_2$). We investigate the impact of channel variability, $\sigma = \gamma_H^{SLB} - \gamma_L^{SLB}$, i.e., the difference between the high gain channel and the low gain one. (For example, the left most point of Fig. 4.4 has $\gamma_H^{SLB} = 10$dB, $\gamma_L^{SLB} = 9$dB, $\sigma = 1$dB and the right most point has $\gamma_H^{SLB} = 18$dB, $\gamma_L^{SLB} = -4.7$dB, $\sigma = 22.7$dB). γ_H^{SLB} and γ_L^{SLB} values are chosen such that with single-lobe beamforming the receivers would achieve the same average rate and hence we can compare relative rate changes as the variability increases.

As shown in Fig. 4.4, both *Greedy* and the *FH-OMB* heuristic perform almost as good as the optimal *Dyn-Prog*. As both receivers have the same channel distribution, differences in receiver state are likely to cancel out over time and maximizing the instantaneous sum rate as *Greedy* does is a good strategy. Only when one receiver is close to finishing and the other receiver is lagging further behind may it be beneficial to favor the lagging receiver instead. Note that the graph also shows 95% confidence intervals but due to the large number of simulation runs they are very small.

For small channel variability ($\sigma < 3$dB), the maximum sum rate is achieved by serving both receivers for any of the channel combinations, hence *Broadcast* and *Greedy* have the same performance. Once the channel variability is increased beyond this point, beamforming only to the

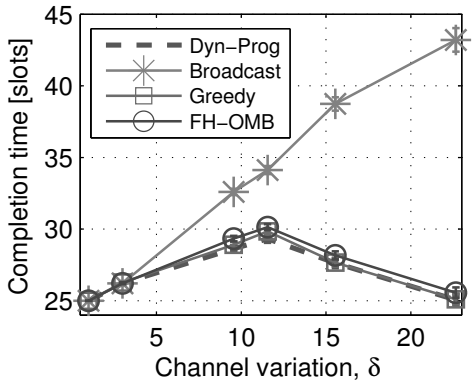


Figure 4.4: Completion time in a homogeneous scenario with increasing channel variability.

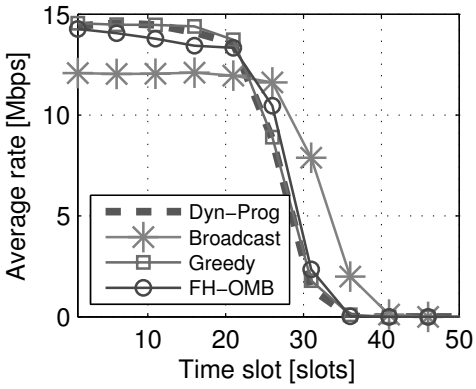


Figure 4.5: Average throughput over time for a homogeneous scenario with channel variability $\sigma = 11.5$dB.

receiver with a good channel when the other receiver has a bad channel (H_1L_2, L_1H_2) provides higher throughput than beamforming to both receivers. Hence, *Broadcast* is unnecessarily conservative by always serving both receivers and its completion time increases substantially as the channels become more variable.

Since in such a homogeneous scenario maximizing sum throughput is almost always the right strategy, *Greedy* even slightly outperforms *FH-OMB* for higher channel variability. Due to this variability, receiver states may differ enough so that *FH-OMB*'s conservative completion time estimate prevents it from opportunistically exploiting good channels as aggressively as *Greedy*. This can be seen in more detail in Fig. 4.5, which shows average system throughput per time slot (averaged over all simulation runs and over both receivers, where receivers that finished have 0 throughput) for the scenario with channel variability $\sigma = 11.5$dB. Throughput of *FH-OMB* starts out the same as that of *Dyn-Prog* and *Greedy*, but drops off slightly once receiver states becomes more heterogeneous and one receiver is close to finishing. Fig. 4.6 shows the completion time estimates for the dynamic programming algorithm (left) and the *FH-OMB* heuristic (right) for the same scenario (i.e., $\sigma = 11.5$dB). *FH-OMB*'s completion time estimate based on average channels underestimates completion time when the channel is more variable, but the relative differences in estimated completion time for the different states for the two algorithms are very similar. *FH-OMB*'s completion time estimation algorithm thus leads to the right beam-forming decisions in most cases. The performance gap is due to the fact that *FH-OMB*'s completion time estimate is slightly less "round" than the true estimate, making it appear more beneficial to stay close to the diagonal where both receivers have the same state.

It is interesting to note that the completion time increases for 1dB $\leq \sigma \leq 11.5$dB and then decreases again. When channel variation is low, both receivers are likely to finish at approximately the same time. The higher σ , the more likely it becomes that one receiver finishes earlier than the other, which increases completion time given by the maximum of the individual completion times. When increasing σ even further, completion times reduce since with a good channel, only

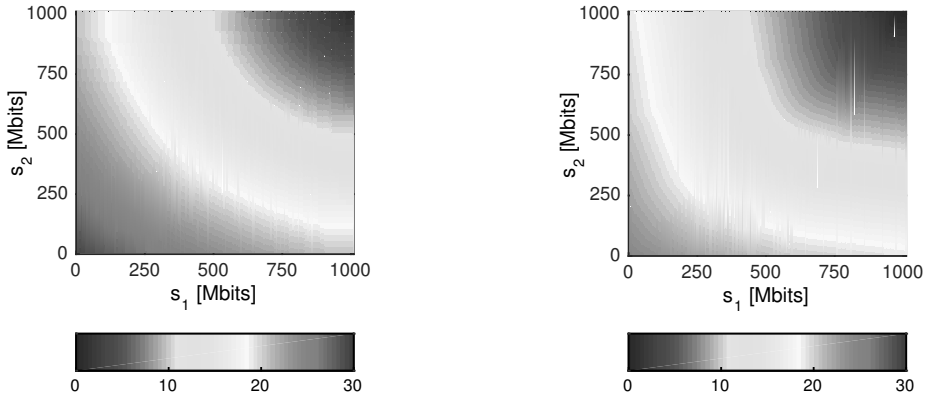


Figure 4.6: Expected completion time for *Dyn-Prog* (left) and *FH-OMB* (right) for $\sigma = 11.5$dB.

very few time slots are needed to complete. There is a significant probability that one of the receivers will finish very early, and the system can then serve the remaining receiver at a higher rate with the corresponding single-lobe beam.

4.5.1.2. Heterogeneous scenario

For the heterogeneous scenario, we fix the $\gamma_{H_1}^{SLB} = 11$dB and $\gamma_{L_1}^{SLB} = -1.4$dB of the first receiver. For the second receiver, we vary $\gamma_{H_2}^{SLB}$ between 11dB and 31dB and $\gamma_{L_2}^{SLB}$ between -1.4dB and 18.6dB, so that the two receivers become more and more heterogeneous as the channel values for the second receiver increase.

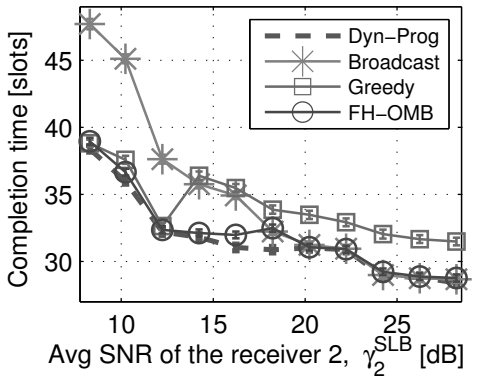


Figure 4.7: Completion time in a heterogeneous scenario with increasing average SNR of the better receiver $\bar{\gamma}_2^{SLB}$ (i.e., receiver 2).

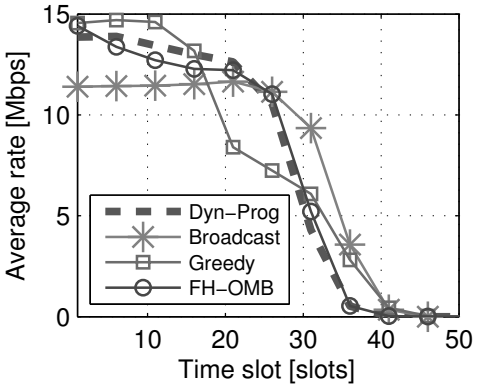


Figure 4.8: Average throughput over time for a heterogeneous scenario $\bar{\gamma}_2^{SLB} = 14.2$dB.

As the $\gamma_{H_2}^{SLB}$ and $\gamma_{L_2}^{SLB}$ increase, completion time decreases for all algorithms. *Greedy* performs close to optimal for the first three data points where receivers are sufficiently homogeneous and the optimum strategy is to beamform to the receiver with high channel gain when one receiver has high channel gain and the other receiver has low channel gain. Here, *Broadcast* is again

too conservative. The jump in *Greedy*'s completion for the next data point is due to the fact that from this point on the good channel of the better receiver is so good that *Greedy* favors the receiver exclusively in that case and only serves both receivers when the good receiver has a low channel gain. In contrast, *Broadcast*'s strategy to balance the rates and forego opportunistic gain becomes closer and closer to optimal as the scenario becomes more heterogeneous and from an average SNR of $\bar{\gamma}_2^{SLB} \geq 17$dB on is the optimal strategy. The weak performance of *Greedy* can be explained from Fig. 4.8, where *Greedy* achieves high throughput until the first receiver finishes at less than approximately 18 time slots. The second receiver is still far from finishing as evidenced by the throughput curve flattening out around 30 time slots. *FH-OMB* performs close but is sub-optimal compared to *Dyn-Prog*, since the expected completion time is slightly inaccurate. The comparison in Fig. 4.9 shows that the expected completion time of *FH-OMB* algorithm (right) is less "round" than that of *Dyn-Prog* (left). Thus *FH-OMB* is more conservative and it sacrifices higher instantaneous rates to ensure that the relative difference in receiver state does not diverge too much.

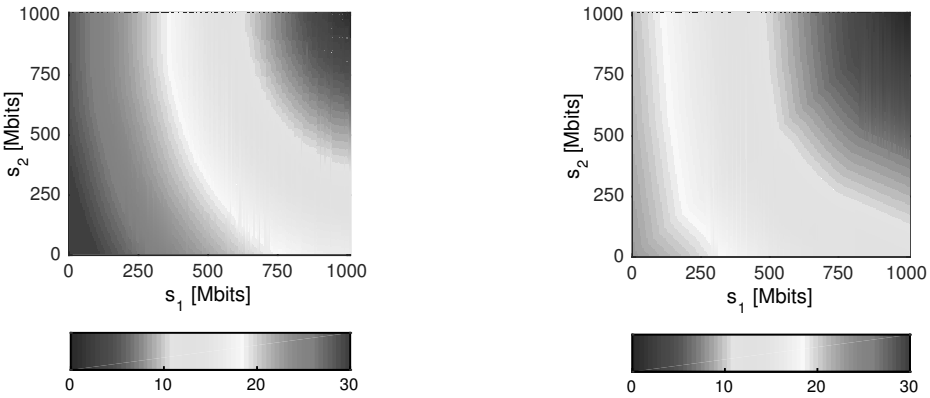


Figure 4.9: Expected completion time for *Dyn-Prog* (left) and *FH-OMB* (right) for $\bar{\gamma}_2^{SLB} = 14.2$dB.

To provide further insights into the behavior of the algorithms we show the state space visits in Fig. 4.10–4.13. As expected *Broadcast* keeps the two receivers very close to the diagonal where both receivers have the same amount of data, and slight deviations from the diagonal are only due to packet loss. *Greedy* in contrast makes quick progress until the second receiver finishes and for the remaining time only has the first receiver to serve. In fact, the steps with which the good receiver makes progress with *Greedy* can clearly be seen in Fig. 4.11. *FH-OMB* serves receivers similar to *Dyn-Prog* early on but then becomes too conservative as the good receiver progresses and beamforms more to the lagging receiver to balance receiver states.

In this section, we show simulation results for a flat multipath Rayleigh fading channel, where the channel does not change within a time slot. The Doppler shift for the Rayleigh channel is set to 10Hz, corresponding to a slow fading channel for receivers moving at walking speed. Receivers are randomly distributed within the coverage area. The BS transmit power is set to 43dBm. With

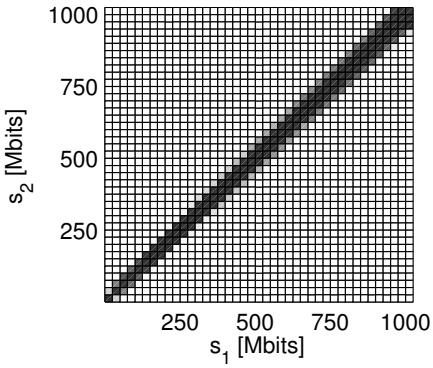


Figure 4.10: State space visits for *Broadcast* at $\bar{\gamma}_2^{SLB} = 14.2$dB.

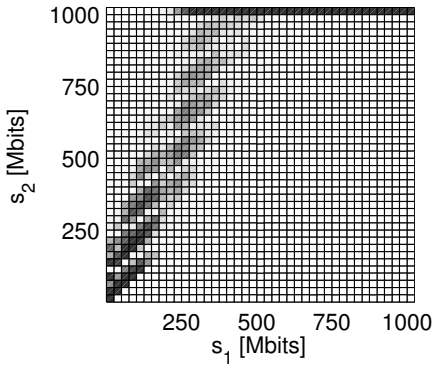


Figure 4.11: State space visits for *Greedy* at $\bar{\gamma}_2^{SLB} = 14.2$dB.

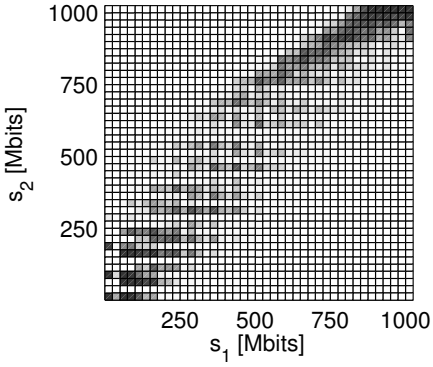


Figure 4.12: State space visits for *Dyn-Prog* at $\bar{\gamma}_2^{SLB} = 14.2$dB.

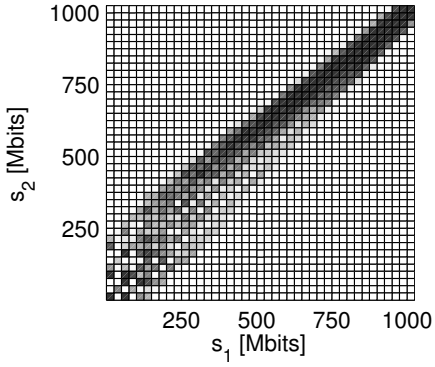


Figure 4.13: State space visits for *FH-OMB* at $\bar{\gamma}_2^{SLB} = 14.2$dB.

this, a cell edge receiver that is 250m from the BS is able to receive a packet with the lowest MCS with an average probability of 30%. The block size B is set to 6400kbits.

We study the impact of increasing the number of receivers N from 2 to 64 with different number of beamforming lobes (i.e., $K = \{2, 4, 8, 16\}$), again for a heterogeneous and a homogeneous scenario. Note that due to the high complexity of *Dyn-Prog*, we only compare the performance of the *FH-OMB* heuristic with that of *Broadcast* and *Greedy*.

4.5.1.3. Random receiver distribution

We first discuss a heterogeneous scenario, where $N = \{2, 4, 8, 16, 32, 64\}$ receivers are randomly distributed within the cell area of radius 250m and for $K = 8$ beamforming lobes. The performance depends significantly on the specific receiver distribution, in particular for smaller numbers of receivers. For up to 16 receivers, *Broadcast* performs almost as good as *FH-OMB* since there is a high probability that there is one receiver with a significantly worse channel than the others (see Fig. 4.14). As the number of receivers increases, a higher number of receivers

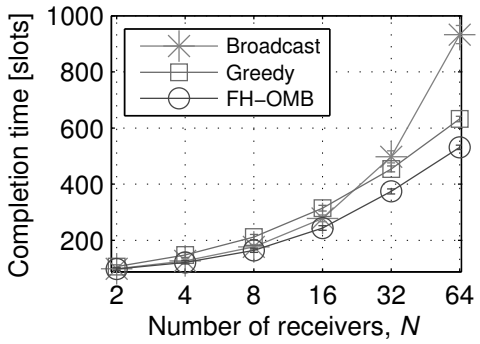


Figure 4.14: Random receiver distribution, $K = 8$.

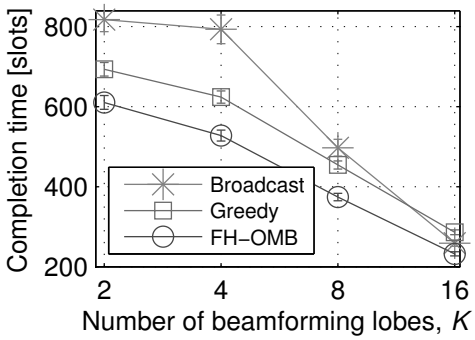


Figure 4.15: Random receiver distribution, $N = 32$.

see similar channel conditions and as in the previous two-channel scenario, the performance of *Broadcast* degrades since it does not exploit opportunistic gain. However, in this heterogeneous scenario this effect occurs mainly for $N > 32$ receivers, where *Broadcast*'s performance is significantly worse than that of *Greedy* and *FH-OMB*. *Greedy* performs worse than *Broadcast* for small N for the same reason as above. The scenario is so small that the receivers are all very heterogeneous. As homogeneity increases for higher network densities, exploiting opportunistic gain becomes more important and *Greedy* outperforms *Broadcast*. *FH-OMB* performs well for all sizes of the receiver set. Its state-based completion time estimation results in the right tradeoff between opportunistic gain and multicasting gain and provides the lowest completion times of all approaches. It consistently outperforms *Greedy* by 9% to 29%. The performance gain over *Broadcast* ranges from 1% to 76%.

Next, we look at the impact of varying K for a fixed $N = 32$. When increasing the number of beamforming lobes K, the array gain of the single lobe beam increases as well. In the specific antenna configuration that is chosen for our simulation, the array gains for $K = \{2, 4, 8, 16\}$ are 1.9, 3.4, 6.6 and 11.4, respectively. (Note that the array gain is not linear in K.) Therefore, as observed from Fig. 4.15, completion time decreases with increasing K with respect to the achievable gain. *FH-OMB* outperforms both *Greedy* and *Broadcast* for all K. However, increasing K has a more significant impact on the completion time of *Broadcast* than on *Greedy* and *FH-OMB*. For low K and a wider beamwidth, *Broadcast* is limited by the receiver with the lowest SNR in each beam. (Also, a significant amount of the radiated energy does not cover any receiver.) As K increases, fewer and fewer receivers are covered by a beam and in the extreme case of a single beam per receiver, *Broadcast* manages to perfectly balance the SNRs at the receivers (i.e, no energy is wasted by having a higher than necessary SNR at any receiver). Hence, *Broadcast*'s performance becomes closer and closer to *FH-OMB*. In contrast, *Greedy* may still beamform to a few receivers with high SNRs so that those finish first, before serving receivers with lower SNRs. In short, in heterogeneous scenarios with sufficient K, *Broadcast* that favors the weaker receivers by multicasting to all the receivers performs better than *Greedy* that capitalizes in maximizing

opportunistic gain.

To shed more light on the behavior of the algorithm, we show the CDF of completion time for the simulation runs for $K = 8$ and with $N = 16$ and $N = 64$ receivers in Fig. 4.16 and Fig. 4.17, respectively. In Fig. 4.16, *Broadcast* and *Greedy* have relatively similar completion time as *FH-OMB* in 10% of the simulation runs. This happens in scenarios where all receivers are distributed quite close to the BS and thus all receivers have a relatively homogeneous good average channel quality. When receivers are distributed sparsely within the cell radius, with high probability they have different average channel qualities. Under this scenario, *Greedy* performs badly since it opportunistically serves the better receivers first and therefore results in higher completion times than both *FH-OMB* and *Broadcast*. Here, *FH-OMB* receivers finish at 465 slots for the worst scenarios, whereas *Greedy* and *Broadcast* both require 570 and 840 slots, respectively. When the number of receivers increases, Fig. 4.17 shows that *Broadcast* no longer has most of its completion time close to *FH-OMB* in most of the simulation runs. In fact, around 20% of *Broadcast*'s completion time is similar to *Greedy* due to the limited number of beamforming lobes ($K = 8$), which leads to low multi-lobe beam's SNR. *Broadcast* is particularly bad in scenarios where many of the receivers are relatively far from the BS (and thus more homogeneous). The worst case completion time of *FH-OMB* is at 675 slots, while *Greedy* and *Broadcast* require 810 and 2400 slots, respectively.

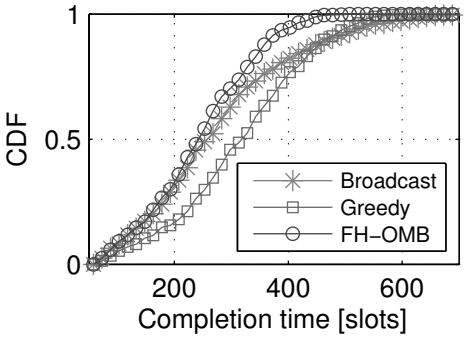


Figure 4.16: CDF of the completion time for random receiver distribution. $N = 16$.

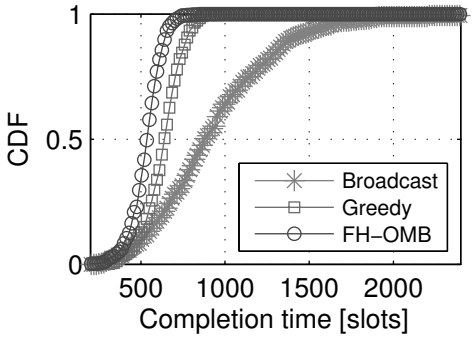


Figure 4.17: CDF of the completion time for random receiver distribution. $N = 64$.

4.5.1.4. Cell edge receiver distribution

In this scenario, receivers are all distributed close to the cell edge in the range of 190m to 220m and thus form a relatively homogeneous group. While such a scenario is less realistic than the one presented in Section 4.5.1.3, it is included to illustrate the performance degradation of the *Broadcast* algorithm in more homogeneous scenarios. Note that this performance is also indicative of the performance in heterogeneous scenarios with very high user densities, where many receivers are at the cell edge (see Fig. 4.14).

Here, the performance differences are much more drastic and *Broadcast* performs worse than

the other schemes already for $N > 8$ (see Fig. 4.18). For 32 receivers, *FH-OMB* outperforms *Broadcast* by 59%. Although maximizing instantaneous throughput is the right strategy for homogeneous scenario, *FH-OMB* still manages to slightly outperform the *Greedy* algorithm by about $1 - 10\%$. Despite the homogeneity of the scenario, the slight differences among the receivers require a more sophisticated mechanism that does take states into account. Similar to the scenario with heterogeneous receiver distribution in Section 4.5.1.3, completion time improves with increasing K (see Fig. 4.19) due to the higher effective SNR for each beam.

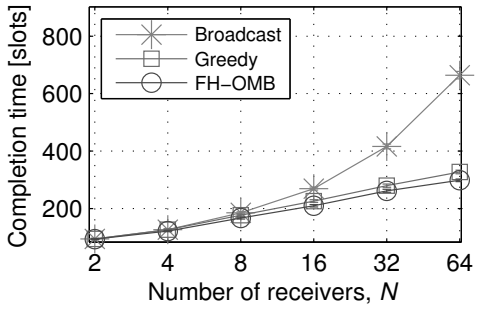


Figure 4.18: Cell edge receiver distribution, $K = 8$.

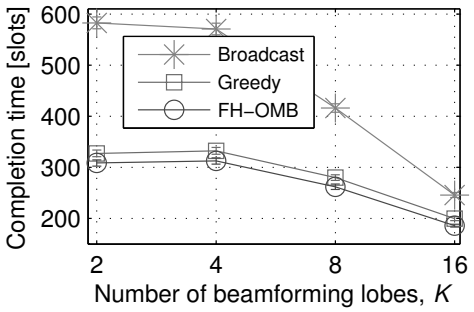


Figure 4.19: Cell edge receiver distribution, $N = 32$.

4.5.2. Fairness of the algorithms.

The fairness criteria that we use is the completion time of the receivers and it is computed based on Jain's fairness index [32]. Given the completion time of each receiver D_i, fairness β is

$$\beta = \frac{\left(\sum_i^N D_i\right)^2}{N \sum_i^N (D_i)^2}. \tag{4.19}$$

Fig. 4.20 and Fig. 4.21 depict the fairness of the schemes when increasing the number of receivers N and beamforming lobes K, respectively. *Broadcast* schedules all receivers and serves them with equal rate and thus it achieves the highest fairness in both settings. Unlike *Broadcast*, *Greedy* maximizes instantaneous rate, allowing receivers with better channels to complete receiving the data block before those with worse channels. This introduces a large difference in the individual completion times between the receivers and yields the lowest fairness among all schemes. *FH-OMB* achieves a much higher fairness than *Greedy* and limits the completion time differences between the receivers. However, *FH-OMB* yields slightly lower fairness than *Broadcast* since it jointly optimizes for instantaneous rate and completion time by serving a smaller subset of receivers than that of *Broadcast* which may cause a greater difference in the completion times.

For a fixed number of beamforming lobes, $K = 8$ (see Fig. 4.20), increasing N increases the density and the diversity of the receivers. For a throughput maximization scheme like *Greedy*, this

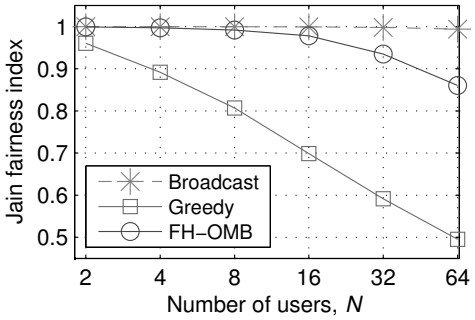


Figure 4.20: Random receiver distribution, $K = 8$.

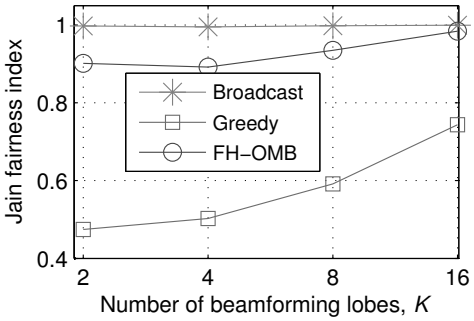


Figure 4.21: Random receiver distribution, $N = 32$.

increases the number of receiver groups to be served. By serving them sequentially (from the best to the worst channel group), *Greedy* creates a difference in completion time between the groups, particularly between the first and the last groups and thus yields a low fairness. *FH-OMB* trades off between multiuser diversity and multicast gain, thus it achieves better fairness than *Greedy*, but with more receivers, the larger variation of completion time among the receivers also reduces fairness.

Increasing the number of beamforming lobes K increases the single lobe array gain and receivers achieve a higher channel gain (see Fig. 4.21). For instance, two receivers (one better and one worse) that are located in the same beam may now be served with two individual beams when K increases. This then increases the chance of both receivers to be served simultaneously without causing throughput loss. Both *Greedy* and *FH-OMB* benefit from this and they thus achieve a significant fairness improvement as the number of beamforming lobes increases.

In summary, although *FH-OMB* sacrifices up to 15% in terms of fairness, but in turn needs half the completion time of *Broadcast* as shown in Fig. 4.19.

4.5.3. Impact of imperfect feedback

To examine the impact of imperfect feedback, we introduce different intervals at which the feedback frames are sent: $\lambda = \{5, 10, 20, 40, 80, 160\}$ ms.[4] Since the coherence time of the multipath Rayleigh fading channel is $\tau = 40$ ms, the actual channel state information and the one reported in the last feedback frame (for simplificity, we call it *last* state information) are correlated for up to τ ms and uncorrelated otherwise. For instance, when $\lambda \geq \tau$ ms, the current and the last state information is only correlated for τ ms after the feedback was received and is uncorrelated for the remaining $(\lambda - \tau)$ ms. Note that, although all receivers send the feedback frame every λ ms, the slot at which the feedback frame is sent is asynchronous among the receivers. [5]

In this section, we also include the corresponding schemes for *Broadcast*, *Greedy*, and

[4] $\lambda = 5$ ms means that the BS receives a feedback frame from a receiver every 5 ms.

[5] For instance, when $\lambda = 160$ ms, the current and the last state information is uncorrelated for $\lambda - \tau = 120$ ms after time t.

FH-OMB in which the estimation algorithms explained in Section 4.4.3 are applied. We call these schemes *eBroadcast*, *eGreedy*, and *eFH-OMB*, respectively. While *eFH-OMB* operates as described in Section 4.4.3, *eBroadcast* and *eGreedy* inherit their own mechanism for their decisions but use the estimated channels as explained in Section 4.4.3 instead of the last state information.

Impact of λ on completion time: Fig. 4.22 depicts the impact of increasing λ on the completion time of the abovementioned schemes (with and without estimation). As λ increases, the correlation between the current and last state information decreases, which causes a scheme to select an inaccurate weight for the beamforming pattern. A too high beam weight wastes resources and a too low one reduces reception probability. This is observed through Fig. 4.22 where increasing λ cause increase in the completion time for all schemes.

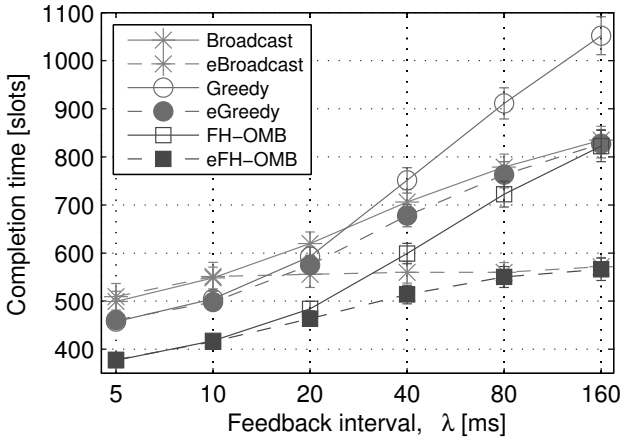


Figure 4.22: Impact of delay on completion time for random receiver distribution, $K = 8$, $N = 32$.

The impact of increasing λ on *Broadcast* is less severe compared to that of *Greedy* and *FH-OMB* because *Broadcast* is conservative. It always beamforms towards all receivers and transmits at a lower rate, thus reducing the magnitude of the error made. While *Broadcast* always schedules the receivers that have not received a complete data block (the subset of the scheduled receivers is fixed), *Greedy* and *FH-OMB* have a higher diversity of receiver subsets. Therefore, inaccurate state information causes *Greedy* and *FH-OMB* to not only make an error in the beam weights (like *Broadcast*) but also in the subset of scheduled receivers. These errors cause a more pronounced impact of λ on the completion time of *Greedy* and *FH-OMB* than on *Broadcast*. Although *FH-OMB* performs similarly to *Broadcast* at $\lambda = 160$ms, it happens for different reasons. When feedback is largely outdated (i.e., the correlation between the actual and last state information is very low) but *FH-OMB* assumes feedback is perfect, this causes errors in the computation of the expectation completion time (see Section 4.4.3 for further explanation). Consequently, *FH-OMB* chooses the wrong subset of receivers to be served and they are served with a wrong MCS, which leads to high completion time for larger λ. *Greedy* performs worst

since it highly depends on the instantaneous channel knowledge to exploit the opportunistic gain. Without instantaneous channel knowledge, a beamforming pattern selected based on the last state information is no longer throughput-optimal.

As shown in Fig. 4.22, with the estimation algorithms detailed in Section 4.4.3, *eFH-OMB*, *eGreedy*, and *eBroadcast* achieve a substantial improvement over their corresponding schemes where estimation is not applied. For $\lambda < 20$ ms, *eFH-OMB* and *eGreedy* outperform *eBroadcast* since they can still exploit the opportunism because the actual and the last state information are correlated. When λ is large, the variation of the actual channel is large and thus the estimated channel approaches the average channel (see Section 4.4.3 and Fig. 4.3 for more details). In this case, opportunistic gain can no longer be exploited and the best strategy is to equally serve all the receivers. Therefore, we observe good performance of *eBroadcast* at $\lambda = 160$ms in Fig. 4.22 since serving at a wrong rate (low MCS) has minimal to no impact on the better receivers within the same beam lobe. *eFH-OMB* performs close to *eBroadcast* since the expected completion time restricts the opportunistic gain (as detailed in Section 4.4.3) and thus *eFH-OMB* is more conservative and schedules more receivers as *eBroadcast* does.

Impact of λ on the number of successful receivers: To further understand the characteristics of the schemes, Fig. 4.23 shows the impact of increasing the feedback interval on the average number of receivers that successfully receive the transmitted data block from the BS (which we termed as *successful receivers*) for all time slots over 100 randomly distributed scenarios. Intuitively, a larger λ causes a higher error, thus reducing the average number of successful receivers as shown in Fig. 4.23. Schemes with estimation generally outperform those without since they improve the accuracy of the state information using the estimation algorithm. It improves the number of successful receivers for *eBroadcast*, *eGreedy*, and *eFH-OMB* by 17.65%, 25.00%, and 83.33%, respectively, compared to the respective schemes without estimation (i.e, *Broadcast*, *Greedy*, and *FH-OMB*). As *FH-OMB* highly depends on the accuracy of state information to make its optimal decision, estimation algorithm is of high relevance. Therefore, we observe *eFH-OMB* (*FH-OMB* with estimation algorithms) achieves the highest improvement over *FH-OMB* as compared to the other schemes when compare against their corresponding scheme without estimation. It is important to note that a higher number of successful receivers do not indicate better completion time. For instance, *eBroadcast* and *Broadcast* achieve a higher number of successful receivers than *eFH-OMB* and *FH-OMB*, respectively, due to the nature of the algorithm which serves all the remaining receivers, but they have a higher completion time due to the low transmission rate. In contrast, *eGreedy* and *Greedy* yield the lowest successful number of receivers since they serve a different group (receivers that maximizes the instantaneous throughput) of receivers sequentially. *eFH-OMB* serves the receivers based on both instantaneous and future progress (completion time) by serving as many receivers as possible without compromising the completion time, thus it achieves a reasonably high number of successful receivers but maintaining a low completion time (see Fig. 4.22).

Impact of λ on the distribution of MCS: As the chosen rate for the receiver served at each

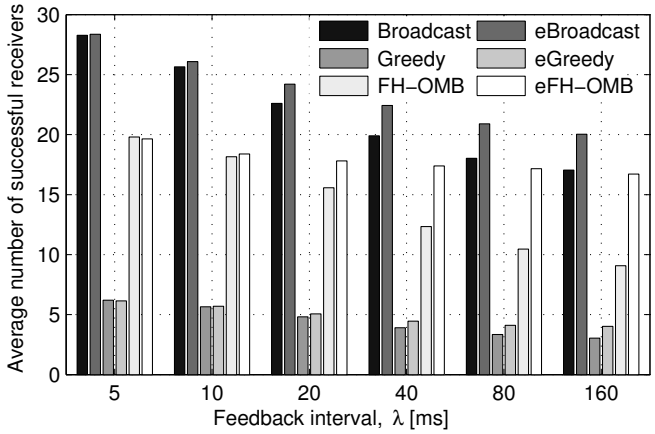


Figure 4.23: Average number of successful receivers for random receiver distribution, $K = 8$, $N = 32$.

slot determines the performance of the algorithm, we take a further look into the distribution of the MCSs used by each scheme for increasing λ as shown in Fig. 4.24. Note that there exists a correlation between the average number of successful receivers and the MCS distribution: when more of the higher MCSs are used (i.e., serving a smaller subset of better receivers), the average number of successful receivers is lower and vice versa. As depicted in Fig. 4.24, *Broadcast* and *eBroadcast* that serve more receivers mostly use the lower MCS and *Greedy* and *eGreedy* uses higher MCSs which serve fewer receivers. In general, the distribution of MCSs is impacted by the remaining receivers in the system as well as the distribution of the estimated channel (for schemes with estimation algorithm).

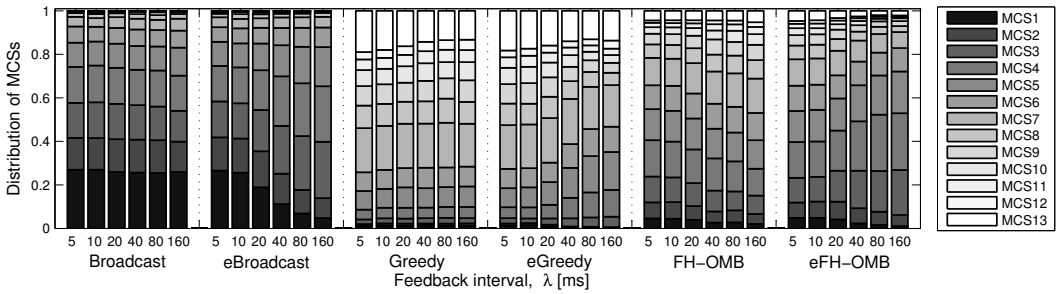


Figure 4.24: Distribution of MCSs for random receiver distribution, $K = 8$, $N = 32$.

For *Broadcast*, we observe a slight increase in the number of the higher MCSs when increasing λ. A higher λ causes an error in the computation of the beam weight, thus some receivers lose the data packet and are served at a later time. At this later time, fewer receivers remain in the system and the BS has to beamform in fewer directions. This leads to higher gains for the beams and thus on average higher MCSs are used. With estimation, *eBroadcast* serves the receiver at the rate close to the average channel rate of the worst receivers at each beam lobe as λ

increases. Therefore, we observe an increasing number of intermediate MCSs (that are suitable for the estimated channels) as λ increases.

For *Greedy*, better receivers leave the system earlier, even for higher λ. Here, a wrong MCS choice has minimal impact on these receivers since they progress quickly once they are successfully served (even if a slightly lower MCS is used). However, the remaining (worse) receivers can only be served using the lower MCSs and wrong usage of MCSs causes more tries to successfully serve the subset of receivers which (based on the last received feedback) maximizes throughput. Since the number of attempts to serve the remaining receivers is higher than that to serve the receivers with excellent channels, more lower MCSs usage is seen from Fig. 4.24 for *Greedy* as λ increases. Although based on the same distribution of estimated channels, *eGreedy* has a different MCS distribution (more higher MCSs) than *eBroadcast* due to the opportunistic nature of the algorithm. The frequency of higher MCSs reduces with increasing λ since the opportunism is limited by the estimated channels.

We also observe an increasing number of higher MCSs for *FH-OMB* as λ increases. For higher λ, the correlation between the actual and the last state information is reduced and *FH-OMB* can no longer compute its tradeoff correctly. The impact is twofold: *(i)* it requires more tries to serve a chosen receiver group, *(ii)* some receivers receive the data block at a much later time. Since *FH-OMB* is neither biased towards the worse nor the better receivers, it uses MCSs that are suitable for the receiver group, which in this case are the MCSs greater than MCS5. Therefore, when more tries are required, an increase in these MCSs is observed. With estimation, the number of intermediate MCSs of *eFH-OMB* increases with λ. This is however not usually due to the increased number of tries, but due to the distribution of the estimated channels (refer Section 4.4.3 for more details) which limits the opportunistic gain. As a result, *eFH-OMB* conservatively serves the receiver as *eBroadcast* does, therefore they have a similar MCS distribution, particularly for $\lambda = 160$ms.

4.6. Conclusion

In this chapter, we studied opportunistic multicast beamforming for the finite horizon problem, where a base station has a fixed amount of erasure-coded data to transmit to multiple receivers. We modeled the problem as a dynamic programming problem to obtain the optimal solution. Due the high complexity of this approach, we designed a heuristic algorithm, *FH-OMB*, that captures the characteristics of the optimal solution and provides a performance that is very close to it. We evaluated *FH-OMB*'s performance both for a discrete channel model as well as multipath Rayleigh fading.

We observed that in the more realistic Rayleigh fading scenario, the performance gains of our *FH-OMB* heuristic are much more pronounced than that in the simple scenarios with two receivers and two channel states. It outperforms other schemes based on maximizing the minimum SNR and broadcasting to all receivers (*Broadcast*), as well as greedily maximizing sum rate

(*Greedy*). Compare to *Broadcast*, *FH-OMB*'s gain increases as the number of receivers increases since *Broadcast* does not exploit opportunistic gain. Also, *FH-OMB* is particularly beneficial over *Greedy* in heterogenous receiver scenarios because it trades off between multi-user diversity and multicast gain results in lower completion times. It improves performance by up to 76% over *Broadcast* and up to 29% over *Greedy* for heterogeneous scenarios with Rayleigh fading. For homogeneous scenarios, these gains are up to 122% and 10%, respectively. Similar (though slightly lower) gains are obtained for the simpler scenarios with a discrete channel model. This chapter additionally addressed the impact of imperfect feedback and estimation algorithms are designed to counter the performance degradation due to this impact. With estimation, our proposed scheme (*eFH-OMB*) outperforms *eGreedy* and *eBroadcast* with estimation by up to 46.14% and 34.62%, respectively.

Part II

Millimeter-Wave Communications

Chapter 5

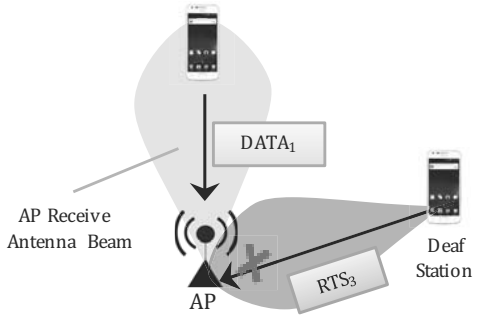
Introduction to Millimeter-Wave Communications

Communication in the mm-Wave band (i.e., 30 to 300 GHz) is a promising next step for the evaluation of high speed WiFi and is considered a key technology for 5G networks. Beside the vast amount of available spectrum. For instance, there is 7 GHz available bandwidth for unlicensed use in the 60 GHz band. In addition, its potential for high spatial reuse of frequency resources makes it very attractive for future dense networks.

mm-Wave systems use directional antennas for communication to cope with high attenuation and transmission losses. At mm-Wave frequencies electronically steerable phased antenna arrays can be implemented with a very small footprint. Thus it is possible to align the beams of transmitter and receiver accurately, which provides the high directional antenna gain to overcome the high attenuation at mm-Wave frequencies. This shift to directional communication raises new challenges for the MAC layer design of WiFi systems. Firstly, this requires accurate alignment of the transmitter and receiver beams. Secondly, the unavailability of reliable omni-directional communication requires significant changes to the system design. Thirdly, stations that are located outside of the transmitter antenna's boresight cannot overhear or sense the transmission, leading to a problem commonly known as the *deafness* problem. Consequently, one of the big drawbacks is that CSMA/CA mechanisms suffer impairments through to directional communication, impacting efficiency and fairness.

Above figures illustrate the problem encountered by a mm-Wave system. The figure on the left depicts an RTS collision caused by two stations being deaf towards each other given their beams directed towards the AP. In the figure on the right, an RTS is lost because the Access Point (AP) focuses its receive beam away from the deaf station, towards a data transmission. The IEEE 802.11ad amendment for 60 GHz WiFi [30] addresses these challenges by defining a hybrid medium access mechanism. This includes polling and Time Division Multiple Access (TDMA) as well as the well known IEEE 802.11 CSMA/CA [46]. However, polling introduces overhead at all stations independent of whether they are communicating or not. Also, TDMA requires ad-

RTS collision due directional transmit focus on AP.



Missed RTS due to receive sector misalignment.

ditional coordination and communication to determine the schedule. In contrast, CSMA/CA is very efficient in handling unpredictable burst traffic, such as the request/response type traffic generated by web browsing. Hence, adaptation of the CSMA/CA scheme to the constraints imposed by mm-Wave communication is of high interest. Hereby, especially frequencies at the 60 GHz band have recently come into focus as the band was released for unlicensed use.

In its basic form, CSMA/CA suffers from several drawbacks when applied to directional communication systems, the main challenge being incomplete carrier sensing. Further, erratic deferral behavior and increased collisions lead to reduced fairness in terms of channel access. While long term fairness is still achieved, a group of active nodes may dominate the channel whereas long time inactive users have a low probability of successfully accessing the medium.

The following chapters (Chapter 6 and Chapter 7) aim to address the aforementioned issues and propose a MAC-layer protocols and scheduling algorithm, respectively.

5.1. Background: IEEE 802.11ad Millimeter-Wave WiFi

IEEE 802.11ad brings WiFi communication to mm-Wave frequencies. This amendment to the IEEE 802.11 standard was ratified in late 2012 and defines significant changes to 802.11 to adapt it to the new frequency band. In this section we highlight relevant changes necessary to enable communication on mm-Wave frequencies.

5.1.1. Beamforming training

Due to the increased attenuation at the mm-Wave band, directional signal transmission is required to achieve reasonable communication ranges of up to a few tens of meters. This requires adjusting the antenna beam direction to focus the signal energy on a direct or a strongly reflected path between two transceivers. To this aim, IEEE 802.11ad proposes a two level beam training protocol to set up a directional link between two stations [46]. Omni-directional communication range at mm-Wave frequencies is very limited and the antenna gain resulting from beamforming

is required at least at either the receiver or the sender (and ideally at both). Thus, an initial connection is established by applying a directional sector sweep from a station, while the other pairing station uses a quasi-omni-directional antenna pattern. This happens during the Sector Level Sweep (SLS) phase which is followed by a Beam Refinement Phase (BRP). During the BRP, directional antenna configurations on the receiver and transmitter side are tested against each other to also configure a receive antenna pattern and fine tune the previously selected antenna configurations.

5.1.2. Hybrid medium access control (MAC)

To cope with the challenges imposed by directional medium access, IEEE 802.11ad defines a hybrid MAC layer. This ensures that also delay critical applications as for example wireless displays can be supported. Besides the WiFi-typical CSMA/CA access scheme, a TDMA scheme as well as polling based access are supported. The AP schedules contention free service periods (SP) dedicated to a specific pair of stations and Contention Based Access Periods (CBAP) throughout the data transmission phase. In this chapter, we primarily focus on channel access in CBAPs, which follows CSMA/CA as described in the following section. Medium access between beacons can follow multiple access schemes, but an AP can also dedicate the entire data transmission time to contention based access. Contention based access is relatively simple, well understood, and particularly suitable for bursty and unpredictable traffic, where the complexity of adaptive TDMA scheduling and the overhead of polling are undesirable.

5.1.3. Contention based access

The IEEE 802.11ad contention based access follows a standard CSMA/CA approach. In general, stations start (or resume) a random back off counter a DCF Inter-Frame Space (DIFS) interval after the end of the acknowledgment of a data frame. The backoff counter decreases at each slot which equals 5μs. Once a backoff counter reaches zero, the corresponding station wins a Transmit Opportunity (TXOP), where it can exclusively transmit one or more frames to another station. Stations overhearing an ongoing frame, track its duration to maintain a Network Allocation Vector (NAV) and defer from decreasing their backoff counter. This process is also referred to as virtual carrier sensing. If a station senses the channel to be busy (either by virtual or physical sensing) or a frame transmission fails, it doubles its contention window until the maximum contention window size of 1023 slots is reached. After successfully accessing the channel, a station resets its contention window to the minimum of 15 slots.

IEEE 802.11ad adapts its CSMA/CA mechanism to directional medium usage. Idle stations generally listen with a quasi-omni-directional receive pattern as the direction of the next incoming transmission is unknown. Thus, directional antenna gain is only achieved at the transmitter side, requiring a robust modulation coding scheme. Therefore, the first frames exchanged are a directional RTS/CTS pair at the most robust coding modulation scheme. The RTS/CTS ex-

change further increases the chance of main interferers within the antenna's boresight to sense the ongoing transmission and refrain from interfering.

Further, IEEE 802.11ad enables spatial sharing during CBAPs by modifying the deferral behavior. Instead of deferring whenever a frame is overheard, a station might still initiate transmission when the receiver is known to be idle. This leads to multiple transmissions at the same time. However, in a pure uplink scenario only one transmission between a station and the AP takes place at the same time.

5.1.4. Fast session transfer (FST)

Wireless links at mm-Wave frequencies are less robust compared to communication at legacy WiFi frequencies at 2.4 or 5 GHz. The main reason is the severe attenuation by blockage that easily interrupts mm-Wave links. This can happen due to one of the transceivers moving behind an obstacle, or, in case of stationary transmitters, due to human blockage [66]. A further drawback for mm-Wave links is the limited range, caused by increased free space attenuation and transmit power regulations.

To overcome these impairments and provide a user experience better than or at least comparable to legacy IEEE 802.11 networks, IEEE 802.11ad supports multi-band communication in the form of a Fast Session Transfer (FST) protocol. The FST mechanism allows a pair of devices to connect over multiple network interfaces at different frequency bands or channels. This is realized either by presenting two different interfaces to the higher protocol layers or in a transparent way, using a single interface with the same MAC address for both physical links. The FST mechanism switches communication seamlessly between high throughput mm-Wave bands and more reliable communication at lower bands for range extension and robustness, i.e., it switches to the lower band whenever the mm-Wave link breaks. *Also simultaneous usage of multiple interfaces is supported by the FST protocol.*

Chapter 6

Multi-Band IEEE802.11ad Millimeter-Wave Networks

6.1. Introduction

Many works have addressed deafness in Wireless Local Area Network (WLAN) and Wireless Personal Area Network (WPAN) presented in [76], however, most solutions are designed for lower frequency communication where omni-directional transmission and reception is feasible and can be used for coordination purposes. At mm-Wave frequencies, however, increased attenuation requires directional antennas at least at one side of a communication link. The most suitable adaptations of CSMA/CA for mm-Wave frequencies are proposed by the IEEE 802.11ad amendment [30] and work by Gong *et al.* [23]. IEEE 802.11ad modifies the CSMA/CA mechanism to protect a data exchange between two nodes with a directional RTS/CTS exchange, which prevents stations with an antenna beam aligned with the transmissions from creating interference. However, as messages are likely to not be overheard by deaf nodes with antenna beams in other directions, these do not defer during ongoing transmissions but unsuccessfully try to access the channel and then excessively increase their contention window. This leads to a fairness problem as station that successfully access the channel have a substantially higher chance to subsequently win the contention again.

A different approach is proposed in [23], where CTS messages are broadcasted by a central controller. To achieve sufficient link budget to receive the omni-directional CTS messages, every station by default directs its receive beam towards the AP. Unfortunately, this approach still suffers from colliding directional RTS messages, which lowers the effectiveness of the deferral process and results in reduced fairness.

Efficiency and fair channel access, i.e., low per packet delay and high throughput, are the major factors that determine the user experience in wireless networks. In this chapter, we address the deafness problem, which deteriorates fairness and efficiency for uplink channels, using a multi-band approach. Use of multiple bands in high speed mm-Wave WiFi is common, due to

easily blocked directional links and limited range, requiring a fall back mechanisms to more resilient lower frequency communication. For example, the IEEE 802.11ad standard specifies so-called Fast Session Transfer functionality to transition between multiple bands [30]. As a consequence, we expect most devices following IEEE 802.11ad and other upcoming mm-Wave WiFi standards to be compatible with our approach.

Our solution maintains high fairness through a combination of 60 GHz communication with control messages on legacy WiFi frequencies. When contending for the channel, stations exchange omni-directional RTS/CTS messages on the lower frequency band to set up a data transmission. The source and destination stations then exchange data frames on the 60 GHz band.

The advantage of our approach is twofold. First, control message exchange on frequencies of 2.4 or 5 GHz is highly reliable over the typically short IEEE 802.11ad communication distances. Thus, every station overhears the control messages, and can correctly defer, avoiding the IEEE 802.11ad unfairness problem. Second, by parallelizing control and data transmission, we free resources on the 60 GHz band for high speed data transmission. Also note that, in contrast to lower frequency networks, RTS/CTS control messages are used by default on the 60 GHz band. Those are no longer necessary with our approach. Thus, the dual frequency approach enhances throughput and MAC efficiency.

The contributions of this chapter are summarized as follows:

1. We analyze the deafness problem in 60 GHz CSMA/CA networks and propose a dual-band solution that couples interfaces on the 60 GHz band with legacy WiFi frequencies.

2. Our mechanism shifts control messages onto a legacy IEEE 802.11 channel with lower bandwidth, freeing up channel time for high throughput transmissions on 60 GHz. By this, we achieve a throughput increase of up to 65.3% over *IEEE 802.11ad* CSMA/CA

3. By exploiting omni-directional transmissions on legacy WiFi frequencies we solve the deafness problem and increase MAC fairness by up to 42.8% compared to *IEEE 802.11ad*.

To the best of our knowledge, this work is the first to address the fairness problem of CSMA/CA in directional mm-Wave networks.

This chapter is organized as follows. In Section 6.2 we discuss past proposals to solve the deafness problem. Section 6.3 gives describe the impairments towards CSMA/CA due to the deafness problem caused by directional transmission. Our dual-band CSMA/CA solution is described in Section 6.4. Section 6.5 provides the channel and transmission model we use for the performance evaluation. Simulation results are shown in Section 6.6 and Section 6.7 concludes the chapter.

6.2. Related Work

Directional antennas are also used in *microwave communication*, e.g., the 2.4 GHz and 5 GHz frequency bands, to improve throughput and reduce interference. Also here, the di-

rectionality of the communication can cause deafness. Solutions to this problem are discussed in [2, 24, 33, 34, 45, 51, 68, 72]. [33] and [45] solve the deafness problem for CSMA/CA systems by omni-directionally transmitting control messages (i.e., RTS and CTS). However, by doing so, all overhearing nodes will defer their transmission, preventing spatial reuse which is particularly important in a directional transmission system. In the same vein, Takai *et al.* [68] propose directional frame transmission while listening omni-directionally. This only partially solves the deafness problem as frames might not reach all receivers due to directionality of the transmission. Korakis *et al.* [34] propose a directional system that emulates omni-directional RTS transmission by sweeping transmit directions, i.e., transmitting an RTS in each possible direction. While this technique effectively deals with deaf stations, it has high overhead, especially for systems with many narrow directional sectors. The works in [2, 24] address the deafness problem by omni-directionally transmitting control messages but they require additional mechanisms to prevent interfering transmissions. Arora *et al.* [2] use separate channels to reduce collisions by adjusting the transmit power such that interference at the receiver is avoided. In [24], an additional GPS receiver is used to provide location information to create a coordination map to avoid interference.

From the above-mentioned works, it becomes clear that even for the microwave band mitigating the deafness problem of directional transmission is not easy, but the availability of omni-directional communication helps significantly. Due to increased attenuation, in general this is not feasible in the 60 GHz band. Given the currently achievable receiver sensitivity and the regulatory transmit power limitations of IEEE 802.11ad, the gain resulting from a directional antenna is needed at least at one side of the wireless link. Therefore, the aforementioned methods that rely on fully omni-directional communication usually do not work for 60 GHz networks.

Only few works address the deafness problem while taking this additional challenge into consideration. Gong *et al.* [22, 23] use the Personal Network Coordinator (PNC) in a WPAN (i.e., similar to an AP in WLAN) to coordinate each transmission. Instead of exchanging RTS and CTS messages between the communicating devices, the source device transmits a directional RTS to the PNC. In response, the PNC broadcast the CTS message omni-directionally. All devices focus their receive antenna in the direction of the PNC and can thus receive the transmission. In case of an uplink channel, this results in a directional RTS transmitted by the station and the PNC broadcasting an omni-directional CTS. This technique partially solves the deafness problem (RTS messages can still collide), but may create a bottleneck at the central PNC. In addition, this technique also prevent spatial reuse since a busy PNC cannot coordinate new transmissions. [44, 64, 65] highlight the importance of solving the deafness problem to avoid collisions in directional 60 GHz networks. However, their approach involves a learning mechanism that resorts to a TDMA-like scheduling. The solution is thus not suitable for the contention based access systems we focus on in this chapter.

The IEEE 802.11ad amendment itself proposes a directional MAC layer mechanism but does not address the deafness problem that leads to a critical fairness issue. We discuss the details of this mechanism in and Section 6.3. Similar to our approach, [47, 48, 52] exploit the coexistence

of a microwave interface. However, there the dual-band approach is used for neighbor discovery, as a fall-back mechanism for range extension, and for beam forming training optimization rather than solving the deafness problem.

6.3. Fairness Impairments in Directional CSMA/CA

This section describes the impact of deafness on the two most relevant CSMA/CA mechanisms for directional networks, IEEE 802.11ad and the centralized approaches proposed in [22, 23].

6.3.1. IEEE 802.11ad CSMA/CA

The deafness problem entails that in many cases other stations will neither overhear frames nor sense that the carrier is busy when two stations are communicating. This causes two important performance impairments described in detail below.

Excessive Backoff. Due to limited (or lacking) carrier sensing, the frame collision probability during contention increases. Especially in dense networks, this significantly increases the average contention window. Furthermore, as frames (including RTS/CTS exchanges) are transmitted directionally, stations outside the transmit beam will not overhear the ongoing communications and thus will not defer. Instead, a deaf station may try to transmit to an already communicating station, which has its receive antenna beam steered into another direction. While may not disrupt the ongoing communication, the deaf station will assume a failed transmission and increase its contention window. Fig. 6.1 depicts this excessive backoff problem. Stations two and three initially collide with their RTS messages, resulting in an increased contention windows compared to station 1. Both have reduced chances to 'hit' the following contention windows. Their next RTS messages will be lost as the AP directs its receive beam away from them due to an ongoing data transmission. Stations two and three further increase their backoff and station one dominates the medium access.

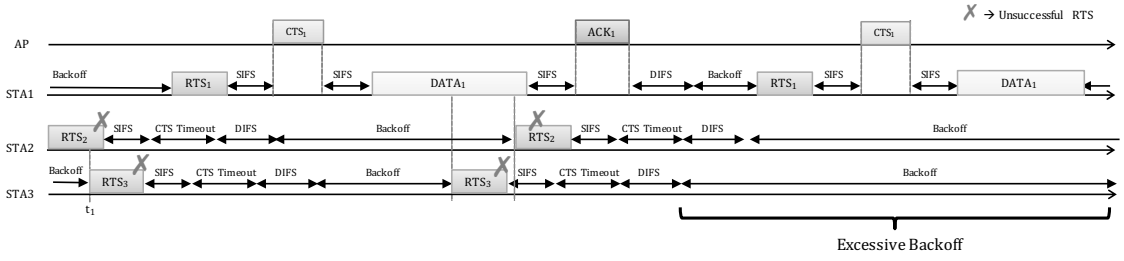


Figure 6.1: Excessive backoff behavior of CSMA/CA in IEEE802.11ad

Unbalanced Contention. Stations suffering excessive back-off have a low probability to win contention, which favors recently active stations with smaller contention windows. While this effect is also present in conventional IEEE 802.11, in IEEE 802.11ad it is vastly exacerbated by

the fact that contending stations do not know when an ongoing frame ends, i.e., when to resume contention. This in turn increases the probability that the same active station transmits a high number of frames consecutively, before other stations happen to win the contention for medium access. This effect primarily impacts short term fairness and over longer time scales, the identity of the active station changes sufficiently often to achieve some level of fairness.

6.3.2. Centralized CSMA/CA

The centralized CSMA/CA schemes [22, 23] partially solves the aforementioned problems through an omni-directional CTS sent by the AP.

However, an omni-directional CTS only ensures correct deferral of overhearing stations in the subsequent data transmission phase. The increased RTS collisions rate due to lack of carrier sensing and the resulting long backoff times remain.

As the duration of an RTS comprises 2 slots, chances for collision are very high especially at network initialization when stations use the minimum contention window of 15 slots. Even for moderate network densities, it is not uncommon that multiple RTS messages collide, resulting in more than two stations increasing their backoff windows in the same contention phase. Interestingly, here IEEE 802.11ad benefits from a significantly lower RTS to RTS collision rate, since RTS messages are often uselessly sent during an ongoing data transmission rather than during the contention phase.

The high RTS collision probability may lead to a rapid increase of the contention windows after network initialization. At the same time, in a highly loaded network, the contention period is relatively small (drawn from the 15 slot minimum contention window). As all stations freeze their contention timer during deferral, a high back off counter needs a significant amount of contention periods to reach zero. This again gives an unfair advantage to currently active nodes with small contention windows to win channel contention. As for networks with deaf stations, it is likely that a small set of active nodes alternately uses the channel, while other stations remain in long periods of repeated maximum backoff. The effect is shown in Fig. 6.2.

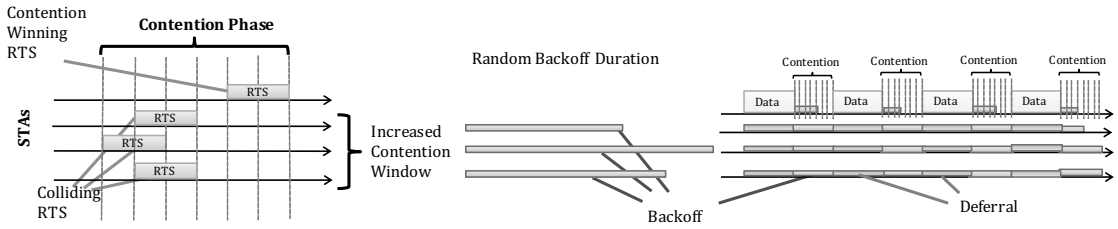


Figure 6.2: Excessive deferral with colliding RTS messages in CSMA/CA with broadcast CTS

Due to an overlapping RTS collision, three stations will increase their contention windows. During the following data transmission the collided stations correctly defer. As a result, their large backoff time is reduced only slowly over the coming short contention intervals (the active station

draws its backoff time from the minimum contention window). The active station thus dominates the channel access for a long period. This effect is more pronounced than for IEEE 802.11ad, as with IEEE 802.11ad even stations with a small contention window are likely to uselessly transmit an RTS during an ongoing data phase and then increase their contention window.

6.4. Dual-Band CSMA/CA

In this section, we propose a dual-band CSMA/CA scheme that mitigates the deafness problem for uplink communication while achieving low overhead and high fairness. Our approach leverages omni-directional transmissions on legacy WiFi frequencies to coordinate high throughput transmissions in the 60 GHz band. As control messages are received by all stations, our approach ensures correct deferral behavior and a low frame collision probability.

We assume an IEEE 802.11ad compatible transceiver design and an infrastructure based network architecture with AP. Further, we require all stations to be able to communicate over a 60 GHz interface as well as over a legacy WiFi interface. This type of transceiver architecture is very likely, as IEEE 802.11ad makes use of a multi-band fast session transfer (compare Chapter 5.1.4) for range extension and seamless failover in case of link breaks. Thus, we expect typical IEEE 802.11ad devices to be compliant with the requirements of our dual-band CSMA/CA scheme. For simplicity we omit details about beam training on the directional 60 GHz interface and assume pre-trained directional links for all stations to the AP. In general, this assumption is satisfied by the association beam training process described in Chapter 5.1.1. Dual-band CSMA/CA access can then be enabled as an addition to the IEEE 802.11ad hybrid MAC architecture of Chapter 5.1.2.

6.4.1. Dual-Band CSMA/CA protocol

Our dual band CSMA/CA protocol follows the random backoff and deferral mechanism, as well as RTS/CTS exchanges as defined for IEEE 802.11ad (compare Chapter 5.1.3). However, the contention mechanism together with the RTS/CTS exchange occur on omni-directional legacy WiFi bands. The IEEE 802.11ad interface of the dual-band devices is exclusively used for data transmission (and acknowledgments).

When applying the contention mechanism on legacy WiFi interfaces, only one message can be exchanged at a time. Thus, for our approach, it is essential to have data frame sizes that exceed the duration of the RTS/CTS exchange consumes on the lower frequency. Otherwise, the data transmission would be delayed by the exchange of control frames. This is especially important considering that lower frequency IEEE 802.11 has longer frame duration of RTS/CTS control messages compared to the 60 GHz band. In addition, since the duration of a data frame is known, it is possible to use in-band RTS/CTS as in conventional IEEE 802.11ad for small data frames for which dual-band RTS/CTS creates too much overhead.

When frames are transmitted according to the dual-band mechanism, stations need to sense the lower frequency band to be idle for at least a DIFS time before starting the contention mechanism. The transmitted RTS frame will reference the end of the latest known transmission on the 60 GHz band plus a Short Inter-Frame Space (SIFS) interval. Receiving RTS and CTS messages omni-directionally ensures that the latest transmission is known to everybody. Note that the RTS/CTS exchange can already occur during the transmission of some previous data frame on the 60 GHz band to avoid unnecessary delay.

Fig. 6.3 shows an example frame flow for our dual-band approach. Three stations (distinguishable by the subscript in frame descriptions) intend to transmit a frame to the AP at the same time. In contrast to deaf IEEE 802.11ad CSMA/CA, backoff happens on the legacy frequency band and the RTS/CTS messages are overheard. As can be seen from frames $Data_1$ and RTS_2, data frame transmission and backoff procedure happen in parallel on two bands.

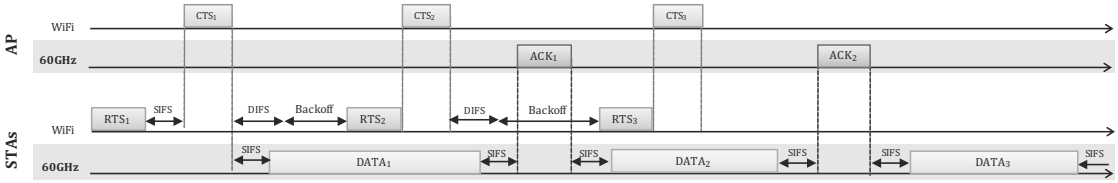


Figure 6.3: Channel access mechanism of the Dual-band approach.

6.4.2. Improvements to millimeter-wave CSMA/CA

A fundamental advantage of our dual-band approach is that all stations can overhear the RTS/CTS exchange on lower omni-directional frequencies, thus solving the deafness problem. As a result, dual-band CSMA/CA does not suffer from the impairments described in Section 6.3. Our approach achieves a flawless deferral behavior, avoids excessive contention windows, and increases the fairness of medium access.

A second benefit results from the parallelization of RTS/CTS exchanges and data transmissions on two separate frequency bands. This removes idle time and RTS/CTS transmissions from the 60 GHz frequency band. As RTS/CTS exchanges are transmitted with the most robust and thus lowest rate coding and modulation scheme, inefficient use of the 60 GHz channel is avoided. Instead, all available channel time with only a SIFS interval between data frames (and acknowledgements) can be utilized for very high throughput transmissions on the 60 GHz channel. Sacrificing transmission time on the legacy WiFi band for control traffic improves efficiency as the band supports much lower transmit rates compared to the 60 GHz band.

Finally, frame collisions due to receive and transmit beam pattern disparity is reduced by the proposed dual-band approach. One of the practical challenges for millimeter wave communication is the generation of undistorted and uniform beam patterns [27]. As a consequence, differences in receive and transmit patterns can lead to a device suffering deafness into certain

directions that are still covered by its transmit beam. The result are frame collisions and disruption of the CSMA/CA protocol flow. For the targeted uplink scenario, the dual-band approach resolves this problem as perfect carrier sensing of the RTS/CTS frames on lower frequencies is provided.

6.5. Simulation Models

This section elaborates the simulation models that are used to produce the numerical results in Section 6.6. We consider an indoor wireless network with a single AP and a set of $\mathbf{S}$ non-AP stations that are randomly distributed within a cell area. The total number of stations is $|\mathbf{S}| = N_s$. Let $\{s, d\}$ represent a transmission pair. We denote the source station as s and the destination station as d.

Channel model. The received power at d when receiving from s is

$$\mathrm{P}_r(s,d)(\mathrm{dBm}) = \mathrm{P_t}(\mathrm{dBm}) + \mathrm{G}_s(\mathrm{dBi}) + \mathrm{G}_d(\mathrm{dBi}) - \mathrm{PL}(l_{s,d}), \tag{6.1}$$

where G_s and G_d are the antenna gains at station s and station d, respectively. The path loss $\mathrm{PL}(l_{s,d})$ including oxygen absorption of stations that are $l_{s,d}$ apart is

$$\mathrm{PL}(l_{s,d}) = 20\log_{10}\frac{4\pi l_{s,d}(\mathrm{m})}{\quad} + \alpha l_{s,d} \quad (\mathrm{dB})$$

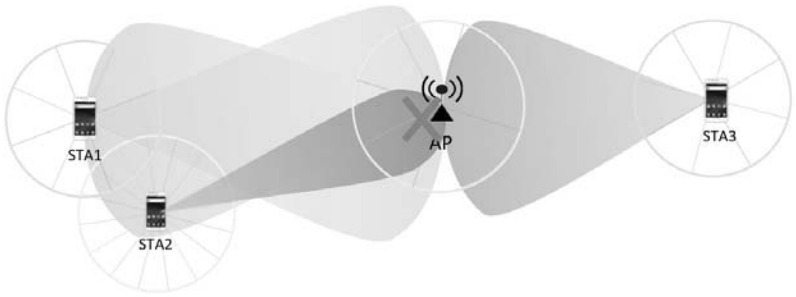


Figure 6.4: Interference in a directional transmission network.

In the presence of interference, the total interference power $\mathrm{P}_{\text{int}}(s, d)$ at the a receiving station d is

$$\mathrm{P}_{\text{int}}(s, d) = \sum_{n \in \{\mathbf{S} \cup \mathrm{AP}, n \neq d, n \neq s\}} \mathrm{P}_r(n, d). \tag{6.2}$$

As mentioned, not all interference causes packet loss. A packet is only lost if the interference signal causes the resulting Signal-to-Interference-plus-Noise Ratio (SINR) degrades below the threshold needed to decode a message coded with a certain coding and modulation scheme. The corresponding thresholds are obtained from [30].

6.6. Results

This section evaluates the performance of our *Dual-band* scheme. We investigate the impact of the number of stations (N_s), time interval τ over which fairness is measured (called "fairness interval"), and frame size (f).

We consider an uplink scenario, where all stations contend for the channel to transmit data frames to the AP. Since all stations are attempting to transmit to the AP, their beams are always directed towards the AP. However, only one station at a time can win a transmit opportunity (TXOP) with the AP. Upon winning a TXOP, a station uses it to transmits a maximally sized data frame that fits the TXOP before recontending for the channel.

Network topology and configuration. In the simulation, stations are randomly distributed within a cell area with a radius of 23m. This range is the same as the maximum range we measured on first generation mm-Wave devices, a Dell 6430u laptop and D5000 docking system with 13° sector width. The fairness intervals τ over which fairness is evaluated are chosen to reflect short term ($\tau = 5$ms) and long term fairness ($\tau = 80$ms), respectively. We also use different frame sizes $f = \{1.5, 15, 30, 45, 60, 75\}$KB to study the impact of different levels of frame aggregation on performance. For data frame transmission at 60 GHz, we consider the 12 single carrier MCSs defined in IEEE 802.11ad [30]. The corresponding transmit rate ranges from 389Mbps to 4620Mbps. Table 6.1 shows the control message transmission rate for 802.11ad and 802.11ac respectively as well as simulation settings.

Table 6.1: Parameters in 60 GHz and 5 GHz frequency bands.

Item	IEEE 802.11ad 60 GHz	IEEE 802.11ac 5 GHz, 80MHz bandwidth
aDIFSTime	13μs	34μs
aSIFSTime	3μs	3μs
aSlotTime	5μs	9μs
MCS0	27.5Mbps	32.5Mbps
aRTSTime = aCTSTime	8.19μs	7.30μs

Performance metrics. Our main performance metrics are throughput, fairness, and delay.

Throughput is the total amount of data bits successfully received at the destination station over the transmission time. To ensure fair comparison, we take the throughput loss on the lower frequency band that our scheme incurs into account. To this aim, we scale the lower frequency band's throughput rate by the amount of channel time not used for RTS/CTS exchanges, and add this to the total rate achieved at mm-Wave frequencies. For non dual-band scheems the full data rate of the lower frequency bands is added, which is fixed to 433Mbps, the maximum rate supported by IEEE 802.11ac without MIMO. Lastly, fairness is computed based on Jain's fairness index [32].

Performance comparison. We compare the performance of the proposed approach against two other schemes. The first approach, *IEEE 802.11ad*, implements the IEEE 802.11ad standard [30] where all messages are transmitted directionally between the source and the destination stations. The second scheme is the one proposed by Gong *et al.* in [23], where an RTS message is transmitted directionally from a source station to the central controller (i.e., AP), which replies with a broadcast CTS message to *all* stations in the system.[1] We denote this centrally coordinated scheme as *Central*.

6.6.1. Homogeneous scenario

Our analysis of the proposed *Dual-band* approach and the comparison to existing methods is threefold. First we evaluate the throughput of the different methods before analyzing fairness and lastly the impact of frame size.

Throughput. Figures 6.5 and 6.6 illustrate the impact of increasing the number of stations N_s on the system throughput and frame collision rate, respectively. Both simulations have a fixed fairness interval of $\tau = 80$ms and frame size of $f = 15$KB. Further, Fig. 6.7 presents the duration for data transmission, MAC overhead, collision time and idle time.

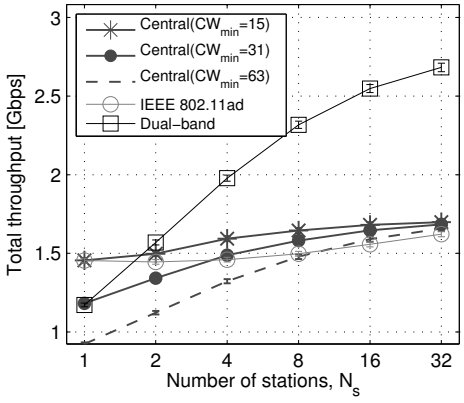


Figure 6.5: Throughput comparison

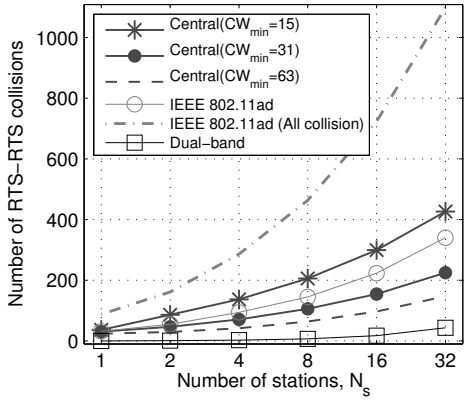


Figure 6.6: RTS-RTS collisions

In general, Fig. 6.6 shows that increasing N_s also increases the probability of RTS frame

[1] Note that this is in contrast to IEEE 802.11ad operation which only permits quasi-omni directional reception, but not transmission at the AP.

collision. While *Central* avoids collisions with data frames due to broadcasted CTS messages by the AP, directional RTS frames during the contention phase still can collide since the directional RTS is not overhead by any other stations. Further, it can be observed that *Central* with higher CW_{min} has a lower number of RTS collisions. The collision rates for *IEEE 802.11ad* are presented twofold. The plain *IEEE 802.11ad* curve only considers collosions between RTS frames, while *IEEE 802.11ad*(All collision) also depicts collisions with CTS and Data frames. The plain RTS-RTS collision rate is comparable to the centralized scheme, however the overall collision rate of IEEE 802.11ad is found to be much higher. In fact, the majority of the collisions in *IEEE 802.11ad* occur due to the transmission of RTS messages from the deaf stations with an ongoing data transmission. *Dual-band* in contrast has the lowest RTS-RTS collision rate since RTS are transmitted omni-directionally and thus stations defer upon overhearing them. Thus, collisions only occur if two RTS messages are transmitted in the exact same slot.

From Fig. 6.5, throughput increases with N_s. This results from the fact that time between two consecutive transmissions is reduced as more stations result in higher probability of a station ending backoff early in the contention phase. Further, increasing the minimum contention window CW_{min} reduces the probability of collision but also causes high throughput loss for small N_s as stations backoff unnecessarily. This effect can be seen for the three different CW_{min} configurations evaluated for throughput and collision rate. According to a detailed analysis on contention based access by Bianchi in [8], increasing CW_{min} bears no significant improvement for systems using the RTS/CTS mechanism. This reflects in the throughput performance in Fig. 6.5 for *Central* where all three configurations different perform equal for $N_s \geq 8$.

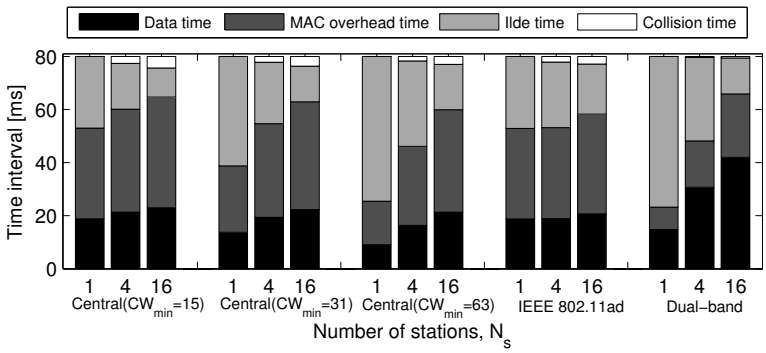


Figure 6.7: Time proportion for data transmission, MAC overhead, idle time and collision time for $N_s = \{1, 4, 16\}$ in the homogeneous sector scenario.

From Fig. 6.7, it can further be seen that the *Central* scheme with a higher CW_{min} incurs more idle time than that with a lower CW_{min}. This shows again, that a higher CW_{min} causes unnecessarily long backoff intervals. In addition, Fig. 6.7 shows that *Dual-band* achieves reduced idle time between frames, which coincides with negative gain for $N_s = 1$. This is due to higher DIFS and slot time on the lower frequency band as shown in Table 6.1. Therefore, the backoff interval time is 80% longer than that for *IEEE 802.11ad* and *Central*. Note that disabling *Dual-*

band in case it is not beneficial is a trivial extension. For higher numbers of stations, the benefits of *Dual-band* and its ability to mitigate the deafness problem outweigh increased inter-frame and slot times.

This analysis shows that Dual-band performs best for $N_s > 1$ and it achieves a throughput gain of up to 65.3% compared to IEEE 802.11ad and 57.9% compared to Central.

Fairness. Fig. 6.8 and Fig. 6.9 show the fairness of the schemes for short term fairnes ($\tau = 5$ms) and long term fairness ($\tau = 80$ms). For both, the frame size is fixed to $f = 15$KB and the number of stations N_s is varied between 1 and 32. Further, Fig. 6.10 and Fig. 6.11 show the histogram of the maximum per frame delay for $N_s = 4$ and $N_s = 16$, respectively from 200 simulation runs. The shown delay duration is the time difference between the current data frame and the next data frame of a station (given that all stations are backlogged). The frame size for these simulations is fixed to $f = 15$KB with a simulation interval of $\tau = 80$ms.

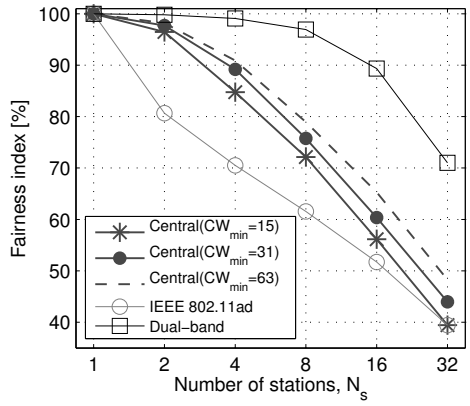


Figure 6.8: Short term fairness.

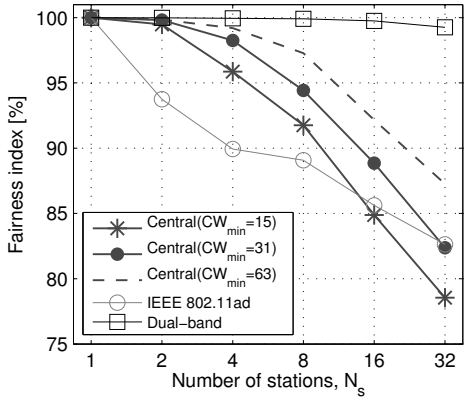


Figure 6.9: Long term fairness.

In general, fairness is found to be higher for smaller N_s. This is because fewer stations have a higher chance to contend for the channel within the simulated fairness interval. Further, Fig. 6.8 and 6.9 show that *Central* and *IEEE 802.11ad* have significantly lower fairness than *Dual-band*. This results from the fact that for the first two schemes, increased RTS collisions lead to longer contention windows. Thus, stations that experience collisions have lower chances of contending for the channel within the fairness interval. In particular, when many stations contend for the channel (larger N_s), most stations excessively increase their backoff interval as the successful contending station repeatedly wins TXOPs with its minimum contention window. The unsuccessful stations defer excessively as their backoff counter needs a long time to reach zero for the contention periods are short (see Fig. 6.2).

Also, *IEEE 802.11ad* suffers from longer idle time as shown in Fig. 6.7, especially when N_s is large. This results from excessively long backoff times due to the deafness problem. However, *IEEE 802.11ad* performs almost as well as *Central*. This is due to the fact that deaf stations continue reducing their backoff timer during ongoing packet transmissions. They thus have a chance to access the channel upon the expiry of the backoff time in a random manner, even when their

initial backoff time was very large. Also, it is found that for both simulation durations, the size of the initial contention windows for *Central* impacts fairness. Smaller minimum windows result in reduced fairness, as nodes recontend after successful transmission with smaller contention times.

For the long term analysis, increased fairness is found for all schemes, since the chances for the stations with larger backoff interval to contend for the channel are higher. Fig. 6.9 further reveals that *IEEE 802.11ad* outperforms *Central* for $N_s \geq 32$. This is because all stations defer access when a CTS is received in *Central*, but only those within the boresight of the AP (that overhear the CTS) will defer backoff in *IEEE 802.11ad*. Thus, deaf stations in *IEEE 802.11ad* continue to reduce their backoff counter instead of deferring. With many deaf stations in a network, thus the chances increase that one of them randomly hits a contention period and interrupts the active station that repeatedly operates on the minimum CW.

These effects also reflect in our frame delay simulation. From Fig. 6.10 for simulations with four stations, it can be seen that *IEEE 802.11ad* has a higher number of frames with a long per frame delay than *Central*. However, for higher number of stations, Fig. 6.11 shows that the distribution of delay duration of *IEEE 802.11ad* and *Central* becomes similar. As explained, this is due to the excessively long deferral time of a station with failed RTS transmission in *Central*, while random RTS transmissions of deaf *IEEE 802.11ad* nodes actually increase fairness for high N_s.

Our fairness analysis reveals that our proposed Dual-band approach achieves significantly higher fairness for arbitrary network sizes because of reduced RTS collisions.

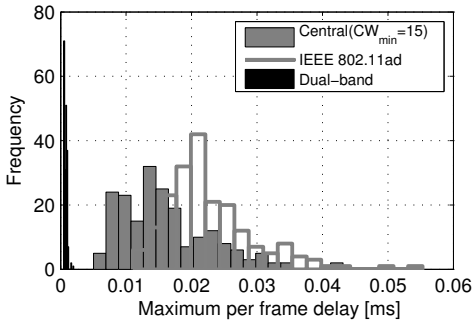


Figure 6.10: Maximum per frame delay for $N_s = 4$.

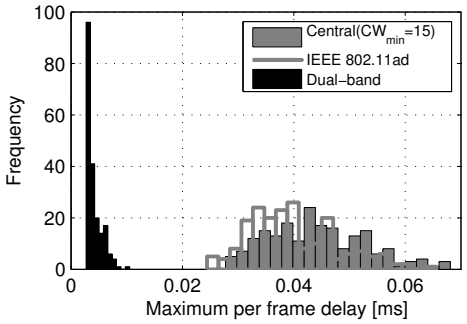


Figure 6.11: Maximum per frame delay $N_s = 16$.

Impact of Frame Size. We further examine the impact of frame size on fairness. Fig. 6.12 and Fig. 6.13 depict the impact of frame size for short term fairness ($\tau = 5$ms) and long term fairness ($\tau = 80$ms), respectively. The number of stations for both simulation runs is fixed to $N_s = 16$. For the fairness of all three schemes, a reduction with increasing frame size is found. For *IEEE 802.11ad*, this results from the fact that larger frame size entails longer transmission time and multiple unanswered RTS of the same station can occur during one data frame. Thus, some stations excessively increase their backoff intervals and have a lower chance to transmit within the simulated fairness interval. Similarly, for *Central* with increasing frame length, stations

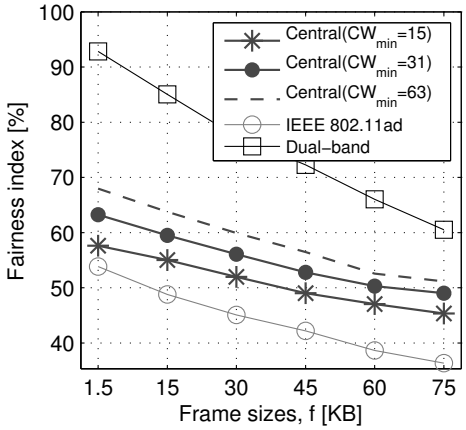


Figure 6.12: Impact of frame size on short term fairness ($\tau = 5$ms)

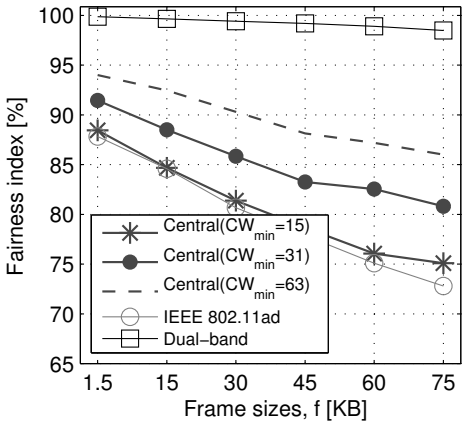


Figure 6.13: Impact of frame size on long term fairness ($\tau = 5$ms)

defer for longer durations and thus some stations may not be able to transmit within the simulated interval. For *Dual-band*, a degradation in fairness with increasing frame size is also found. This results from reduced chances to contend for the channel for longer frame durations as stations are likely to draw a larger (random) backoff times. This effect is less pronounced for shorter fairness intervals in Fig. 6.12. Nevertheless, *Dual-band* performs best for both $\tau = 5$ms and $\tau = 80$ms since it neither has excessive deferral time nor backoff time due to a high RTS collision rate.

Next it is found that for all frame sizes the fairness of the schemes improves for a longer simulation interval $\tau = 80$ms. This is due to the fact that stations that are deferring or backing off for extended periods are more likely to win a TXOP throughout the larger simulation interval. Also, *IEEE 802.11ad* and *Central* perform almost equal in Fig. 6.13 as with longer simulation interval, stations that back off have almost an equal chance to transmit as the stations that defer in *Central*. The fairness differences between *Central* schemes with different minimum contention windows result from the same reasons described in the subsection before. *Overall, it is found that longer frame durations reduce fairness, with the dual-band approach achieving substantial improvements over IEEE 802.11ad and Central. This is because the latter two methods suffer excessive back off and deferral.*

6.6.2. Heterogeneous scenario

We also simulate a heterogeneous scenario where stations have different sector widths, selected at random from the set $\{13°, 15°, 20°, 30°, 40°, 50°, 60°, 90°, 120°\}$. Since the performance is very similar to that shown in Section 6.6.1, we omit the discussion of the results for brevity.

6.7. Conclusion

In this chapter, we address the deafness problem that affects throughput and fairness for directional 60 GHz transmissions with CSMA/CA channel access. To the best of our knowledge, our work is the first to design and analyze a dual frequency system to mitigate the deafness problem. Our approach takes advantage of a coexisting 5 GHz band to coordinate the data transmission at 60 GHz. This is beneficial in two ways. First, due to omni-directional transmission on 5 GHz the deafness problem is solved, preventing fairness impairments. Second, moving robust low rate control messages to 5 GHz allows to use the 60 GHz band exclusively for high throughput data transmission.

We analyze throughput and fairness of our proposed scheme through extensive simulations to compare it against *IEEE 802.11ad* and an alternative scheme that broadcasts CTS messages from a central controller. In terms of fairness, we improve by up to 42.8% over *IEEE 802.11ad* and 34.5% over the centralized scheme. Despite using air time on the 5 GHz band due to broadcasting of control messages, our approach still achieves significant overall throughput gain. Considering both bands jointly, we gain 65.3% and 61.8% over *IEEE 802.11ad*, respectively the centralized scheme.

Chapter 7

Efficient Decentralized Scheduling for 60 GHz Mesh Networks

7.1. Introduction

In view of the significant mobile data traffic growth currently anticipated [10], millimeter-wave (mm-Wave) frequency bands are being explored as a candidate solution to tackle the capacity shortage faced by mobile broadband networks. The very wide (hundreds of MHz to GHz) channels and underutilized spectral resources in these bands open up the possibility of enhancing the capacity of indoor and outdoor wireless deployments and implementing high throughput wireless backhauling. At the same time, mm-Wave bands have high path loss, primarily due to carrier-frequency-dependent attenuation and, secondarily, due to oxygen absorption [79]. To overcome this problem (high path loss), stations employ high gain directional communication, for example through small phased antenna arrays, which allows to confine the emitted energy to narrow beams. This also reduces interference substantially and boosts spatial reuse [69].

Such directional communication, however, introduces *terminal deafness* in the absence of appropriate beam steering and scheduling mechanisms. Therefore medium access solutions previously designed for 802.11 Wireless LANs operating in legacy bands are inappropriate for mm-Wave networks. Beam steering has been addressed in the context of 60 GHz networks that follow the IEEE 802.11ad standard [31], e.g., through out-of-band angle of arrival estimation [47], to reduce throughput degradation associated with transceivers' beams misalignment.

In addition to identifying the right antenna sector or beam direction, scheduling, i.e., *when* to establish a directional link with the intended receiver, is essential to network performance. A simple example scenario is illustrated in Fig. 7.1, where station 2 forwards traffic from 3 and 4 towards station 1. In the absence of appropriate scheduling, station 2 may loose packets of either station 3 or 4 when it communicates with station 1. While simple centralized single-hop scheduling techniques (e.g., the service period based scheduling mechanisms specified by the IEEE 802.11ad standard [31]) may be sufficient for this basic example, they do not scale to more

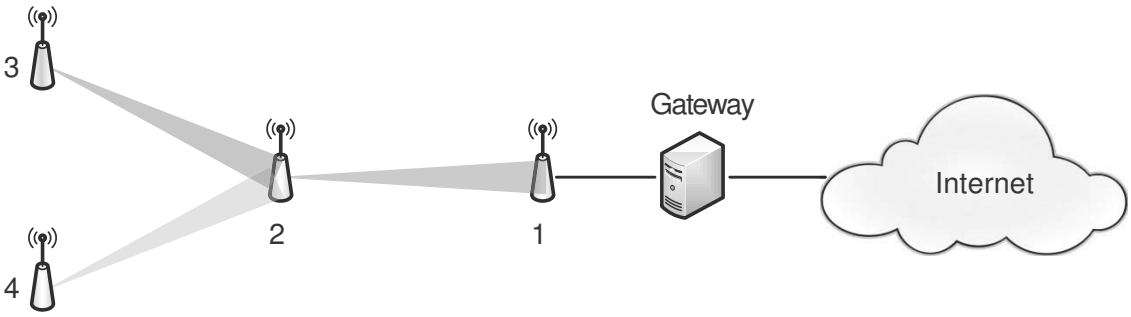


Figure 7.1: Simple example of a multi-hop 60 GHz network. Station 2 forwards traffic originating at 3 and 4, towards the gateway (node 1).

complex multi-hop relay networks. The reason is that stations need to rapidly and precisely decide which neighbor to beam-steer towards for transmission and reception. Finding a single schedule that suits the entire multi-hop network is a complex problem that typically involves global knowledge and coordination.

Medium access control tailored to 60 GHz mesh networks was only considered recently [64], though early efforts fail to capture important practical aspects, including multi-rate operation (due to different link signal-to-noise ratios) and frame aggregation. Precisely, in realistic network settings where links are stable but stations are located at different distances from each other, scheduling over fixed size slots is suboptimal – transmission slots of short duration only allow limited frame aggregation or even require that longer packets are split over multiple slots, whereas long slots are frequently underutilized. In addition, this approach requires alignment of slot boundaries across all stations in the network, which imposes tight synchronization.

In this chapter, we tackle the problem of efficient scheduling in multi-hop 60 GHz networks through a self-organized approach, DLMAC. DLMAC enables stations to *learn in a decentralized fashion* when to trigger conflict-free directional transmissions, without unnecessarily consuming additional channel resources. With this mechanism, stations operate in an unslotted channel that they divide into cycles of fixed length, comprising a number of micro-slots. Stations explore randomly chosen micro-slots within an exponentially increasing access window and upon success, the communicating pair reserves the same time interval for directional packet exchanges in subsequent cycles. After that, the transmitter initiates a backward probing procedure to reduce the idle periods in between adjacent allocations (inter-transmission idle time) and improve efficiency. In addition, we propose a micro-slot binary search enhancement, BinDLMAC, which further reduces the inter-transmission idle periods to boost performance.

We demonstrate by means of extensive simulations that our proposal substantially outperforms recent history-based solutions for mm-Wave mesh networks [64] in multi-rate and variable packet size scenarios, which makes it particularly suitable for indoor high-speed access networks, in-band backhauling and multi-hop relaying. The simulation results show that our approach achieves throughput gains of up to a factor of 8 in single-hop networks and end-to-end throughput gains of up to a factor of 1.6 in multi-hop topologies.

The rest of the chapter is organized as follows. We overview the related work in Section 7.2. We present our proposal in Section 7.3 and evaluate its performance in Section 7.4. Then we conclude the article with some final remarks in Section 7.5.

7.2. Related Work

Recent work provide first-hand practical evidence of the 60 GHz frequency band's capability of multi-Gbps communications [53, 79] and characterize the highly directional mm-Wave wireless links as having some pseudo-wired like characteristics [65]. However, due to the deafness introduced by the highly directional antenna patterns, carrier sensing is infeasible and thus legacy MAC protocols, e.g., the traditional 802.11 operating in the 2.4 and 5 GHz bands, are unsuitable for 60 GHz links.

Learning-based Scheduling in Wireless Networks: Learning was applied previously in the context of traditional wireless networks to achieve TDMA-like scheduling [6, 9, 19, 25, 37, 43, 55, 70, 77]. However, carrier sensing enables these earlier schemes to find collision-free slots and determine the schedule length, which is infeasible in mm-Wave networks without more complex and high overhead exchange of global information. In contrast, our proposal employs decentralized scheduling for multi-hop 60 GHz networks, overcoming terminal deafness without any information exchange.

Multi-hop 802.11 Scheduling Solutions: Scheduling methods designed for the legacy 802.11 multi-hop networks cannot be applied to 60 GHz systems due to fundamental differences that arise with the use of narrow beams. For instance, Choudhury *et al.* propose a directional MAC protocol that employs multi-hop RTS, while CTS, DATA, and ACK are transmitted over a single hop [15]. The approach relies on directional carrier sensing, which cannot be applied on very narrow beams. Laufer *et al.* propose XPRESS, a back-pressure mesh architecture [36], in which a central controller schedules all mesh access points, requiring complex cross-layer information and synchronous operation between the network and link layers. In contrast, DLMAC does not rely on carrier-sensing, tight synchronization, or complex cross-layer interactions.

60 GHz MAC Designs: Given the unique PHY properties of mm-Wave bands, the focus in the design of new MAC protocols is shifted from interference management towards overcoming terminal deafness [44]. In single-hop networks, Chandra *et al.* propose to adapt beam widths in mm-Wave contention-based access [12] to increase throughput, while legacy 2.4/5 GHz bands are employed in [47] to aid mm-Wave technology with beam steering to establish multi-Gbps links. These approaches improve 802.11ad protocol efficiency, but do not address the scheduling problem in the context of deafness.

Chen *et al.* make a step forward and propose a directional cooperative protocol [14], which enables the access point to transform low-SNR single-hop links into multi-hop relayed connections. The solution, however, is centralized and thus has limited scalability in applications such as mm-Wave in-band backhauling. To tackle this problem, Singh *et al.* propose MDMAC, a

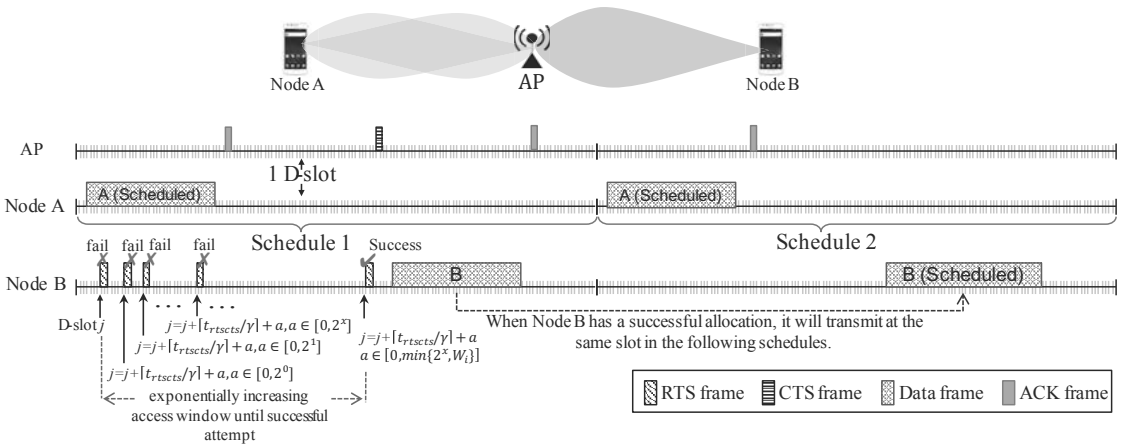


Figure 7.2: Two DLMAC stations accessing the channel: schedule, micro-slots and transmission procedure using an exponentially increasing access window upon failed transmissions.

memory-guided scheduling algorithm that works distributively in directional 60 GHz mesh networks [64]. However, MDMAC addresses scheduling efficiency only to some extent, as the protocol does not capture multi-rate operation and variable packet lengths – it determines a fixed slot size for all transmissions a priori. Further, it involves periodic probabilistic resets of the system state, which further degrades performance. In addition, MDMAC's operation requires synchronization among nodes, which is not trivial in multi-hop topologies. Our proposal tackles these limitations as it does not involve synchronization and is not tied to a fixed slot length. Instead we employ quasi-unslotted access and exploit an effective packing mechanism to improve channel utilization. Consequently, we achieve efficient scheduling in 60 GHz networks under steady channel conditions, but with variable link rates.

7.3. Decentralized Learning MAC Protocol (DLMAC) for 60 GHz Networks

We propose DLMAC, a decentralized learning scheme for scheduling transmissions in 60 GHz networks. Stations running DLMAC decide when to transmit based on the outcome of the previous attempts, with the goal of: *i)* finding conflict-free channel allocations, and *ii)* minimizing inter-transmission idle time. In addition, we specify BinDLMAC, which extends DLMAC through a *Micro-slot Binary Search Procedure* (in Section 7.3.5) to further improve channel utilization.

7.3.1. Protocol overview

Our protocol builds upon the recently approved IEEE 802.11ad standard for 60 GHz networks [31], which mandates that idle nodes listen in quasi-omnidirectional mode and only switch

to directional communication upon a transmission request. This however introduces terminal deafness, i.e. an intended receiver would fail to engage with a transmitter, if already communicating with a different station. To overcome this problem, DLMAC clients independently divide time into cycles of fixed length (schedules) comprised of a number of micro-slots of very small duration and seek to identify non-conflicting sets of micro-slots that can accommodate their transmissions. Stations follow the same cycle length, without requiring to be synchronized and thus the beginning of a cycle can be different for each node.

A node attempting transmission initially picks a set of consecutive micro-slots at random (out of those in the schedule that, to its knowledge, are free) to transmit. If the transmission is successful, the same set of micro-slots will be reserved for future message exchanges in the following schedules at both the transmitter and the receiver. Consequently, both nodes will beam-steer towards each other during the allocated time interval. In case the transmission is unsuccessful, the sender repeats the procedure by choosing at random a different set of micro-slots within an exponentially increasing access window that follows the previous failed transmission attempt.

To improve channel utilization, nodes with established channel allocations probabilistically probe the channel to transmit at an earlier time, with the goal of moving their transmissions closer to other allocations, thereby attempting to cluster packet transmissions together and, thus, prolonging idle intervals to better accommodate future allocations. More specifically, a node will seek to transmit right before its current allocation, such that if the probing is unsuccessful, previously reserved micro-slots can still be used. Figs. 7.2–7.3 summarize DLMAC's operation, which we further detail next.

7.3.2. Scheduling

In contrast to legacy IEEE 802.11, the lack of carrier sensing due to directional communication prevents nodes from inferring the boundaries of other transmissions, which questions the applicability of slotted channel access schemes to 60 GHz networks. Further, mm-Wave protocols where a station maintains synchronization and transmissions are confined to fixed length slots (e.g. [64]) perform sub-optimally with varying packet lengths and PHY bit rates – slots are either underutilized or too small to accommodate large payloads.

To address this issue we propose an asynchronous mechanism whereby nodes divide time into schedules that comprise a fixed number of micro-slots, and select a set of these for communication, as depicted in Fig. 7.2. This approach provides variable-sized allocations to different nodes, allowing DLMAC to adapt better to heterogeneous scenarios with different packet lengths and/or data rates.

We consider the schedule length to be sufficiently long so as to accommodate transmissions in the largest neighborhood and allow for multiple transmissions by the same station in the schedule. Note that a given station may hold multiple allocations within the same schedule, if a suitable set of micro-slots is found for each transmission. In this case, once a node observes that its schedule would not allow for a new station to transmit, it will locally decide to either deallocate one of its

transmissions or a reception (by not sending ACKs).

7.3.3. Reception procedure

A node not participating in any communication listens in quasi-omnidirectional mode, so that it can receive requests for communication from its neighbors. If an RTS is received from a particular neighbor during this phase, the node will first assess whether there is enough time to complete the full exchange of CTS, data packet, and ACK, by checking the time left before its next scheduled transmission or reception. In case the full packet exchange can be completed, it will reply with a CTS and upon reception of the data packet, the node will consider this as a scheduled transmission for the next cycle. This is depicted in Fig. 7.2, where successful allocations in *Schedule 1* are maintained in *Schedule 2*. Right before a scheduled transmission, nodes involved in the communication switch to directional beams and point these towards each other.

7.3.4. Transmission procedure

We now describe how stations transmit (summarized in Algorithm 1). This involves an initial random channel access followed by packing using RTS probing.

7.3.4.1. Initial channel access

A node with a queued packet first sends an RTS in a micro-slot $j \in c(s)$ selected uniformly at random, where $c(s)$ denotes the set of idle micro-slots at schedule $s \in Z^+$. We assume s has sufficient consecutive idle slots to accommodate the transmission[1] (see lines 3–4 in Algorithm 1). If the transmission is successful, i.e., both CTS and ACK are received, the node will consider this attempt as the first successful allocation. The following frames in subsequent schedules are exchanged using the basic access mode (without an RTS/CTS handshake).

If the transmission is unsuccessful, the node infers that the failure may be caused by the receiver being in communication with another station. The node retries in a time slot selected at random from an *exponentially increasing access window* (see Fig. 7.2). To this end, the station draws a random number a in the range $[0, W_i]$, where i is the number of unsuccessful attempts experienced by that packet and W_i is the corresponding access window. The gap between the transmission attempts will be $j + \lceil t_{\text{rtscts}}/\gamma \rceil + a$ micro slots, where $t_{\text{rtscts}} =$ aRTSTime + aSIFSTime + aCTSTimeoutTime, and γ denotes the duration of a micro-slot (lines 13–15). If the attempt is unsuccessful, the station increases the access window and draws randomly a new micro-slot (lines 20–21). This procedure is repeated until a successful transmission occurs.

Note, this design speeds up convergence by backing off rather than waiting for the next schedule. While it would be possible to access the channel more aggressively by continuously sending

[1] Recall that a complete transmission comprises the RTS, CTS, data, and ACK frames, which are separated by short inter-frame spacing times (SIFS).

Algorithm 1 DLMAC - Transmission Procedure

Input:
$j \in c(s), s \in Z^+, W_i, t_{\text{rtscts}}, W_{\max} = 128, p_{\text{rts}} = 1.$
Output: j, W_i
initialize: $i = 0, W_i = 2^i, m = 0.$
choose slot j randomly in $c(s)$
access slot j
if successful **then**
    $p_{\text{rts}} = 1$
    Procedure RTS probing
else
    Procedure Exponential access
end if

procedure EXPONENTIAL ACCESS
    increase the access window: $i = i + 1, W_i = 2^i$
    access range: $a \in [0, \min\{W_i, W_{\max}\}]$
    access at $j + \lceil t_{\text{rtscts}}/\gamma + a \rceil$
    if successful **then**
        $p_{\text{rts}} = 1$
        Procedure RTS probing
    else
        update j; $j = j + \lceil t_{\text{rtscts}}/\gamma + a \rceil$
        Procedure Exponential access
    end if
end procedure

procedure RTS PROBING
    if rand$(1) < p_{\text{rts}}$ **then**
        access at $j - \lceil t_{\text{rtscts}}/\gamma \rceil$
        if successful **then**
            update j; $j = j - \lceil t_{\text{rtscts}}/\gamma \rceil$
            if $m < S/t_{\text{rtscts}}$ **then**
                $m = m + 1, p_{\text{rts}} = 1$
            else
                $m = 0, p_{\text{rts}} = \min\{p_{\text{rts}} p_{\text{red}}, p_{\text{min}}\}$
            end if
        else
            $k = j - \lceil t_{\text{rtscts}}/\gamma \rceil$
            $m = 0, p_{\text{rts}} = \min\{p_{\text{rts}} p_{\text{red}}, p_{\text{min}}\}$
            Procedure Micro-slot binary Search
        end if
    end if
    Procedure RTS probing
end procedure

procedure MICRO-SLOT BINARY SEARCH [BINDLMAC]
    access at $k + \lceil (j - k)/2 \rceil$
    while $\lfloor j - k \rfloor > 0$ **do**
        if successful **then**
            update j; $j = k + \lceil (j - k)/2 \rceil$
        else
            update k; $k = k + \lceil (j - k)/2 \rceil$
            Procedure Micro-slot binary search
        end if
    end while
end procedure

RTSs until transmission is successful, our approach reduces the possibility that transmitter's side lobes may disrupt existing directional links [47].

7.3.4.2. Packing transmissions via RTS probing

To reduce the idle periods between transmissions, nodes try to move their allocations closer to other transmissions in the schedule. To this end, once successful, a node starts *RTS probing* (initially with probability $p_{\text{rts}} = 1$) in subsequent schedules. The station sends an RTS in micro-slot $\lceil t_{\text{rtscts}}/\gamma \rceil$ earlier (line 27 in Algorithm 1), allowing enough time for an RTS/CTS exchange before the original transmission is scheduled. If a CTS is received, the transmission is moved back

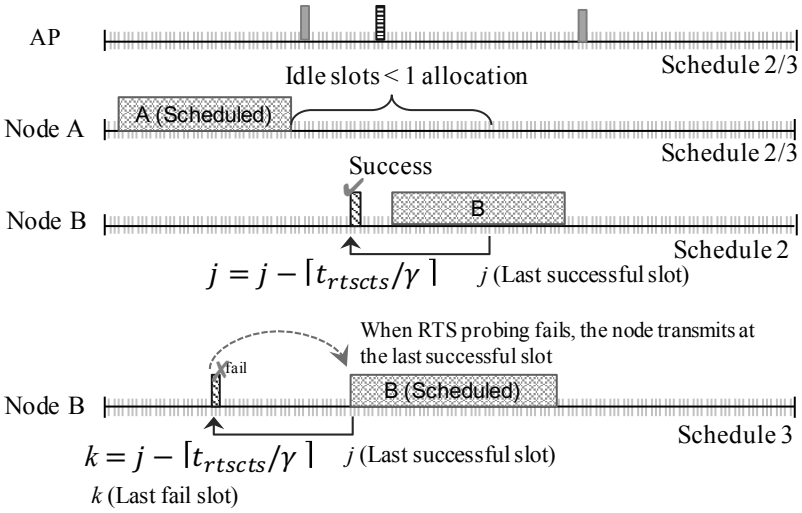


Figure 7.3: RTS probing procedure: attempting to move an allocation $\lceil t_{\text{rtscts}}/\gamma \rceil$ micro-slots earlier in the schedule.

Initially, RTS probing is limited to a maximum of S/t_{rtscts} times, where S denotes the length of the schedule in seconds (line 31). After this, RTS probing is probabilistically used to address potential gaps caused by nodes leaving the network, while limiting the amount of probing when conditions are more stable. For the initial transmissions, we use $p_{\text{rts}} = 1$ (lines 6 and 17). After a failure in RTS probing (line 37) or when reaching the maximum number of attempts (line 33), a station updates p_{rts} to $\max\{p_{\text{rts}} p_{\text{red}}, p_{\min}\}$, where p_{red} is a reduction factor to p_{rts} to gradually lower the RTS probing probability and $p_{\min}$ is a minimum probing probability to ensure the frequency of RTS probing does not become too low. This ensures that when nodes release allocations, arising gaps will be packed quickly. p_{red} and $p_{\min}$ are configurable parameters and we provide suitable values in Section 7.4.

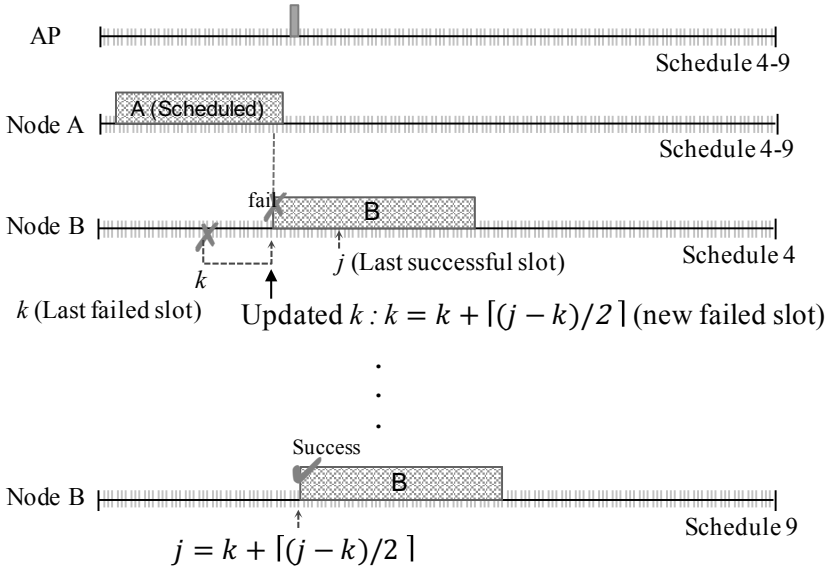


Figure 7.4: Micro-slot binary search phase: attempting to transmit at an earlier slot and cluster allocations.

7.3.5. Micro-slot binary search

Finally, we define a micro-slot binary search mechanism as an extension to DLMAC, referred to as BinDLMAC, to further improve efficiency by minimizing the inter-transmission idle periods. The refinement is motivated by the observation that DLMAC may leave idle time of a duration up to t_{rtscts} between consecutive transmissions. In the *micro-slot binary search* depicted in Fig. 7.4, a node considers j (its currently allocated micro-slot) and $k = j - \lceil t_{\text{rtscts}}/\gamma \rceil$ (the point at which the last *RTS probing* failed). In the next schedule, the node attempts moving its allocated transmission to $k + \lceil t_{\text{rtscts}}/2 \rceil$. Then, upon failure, the node updates k to the new failure point (see line 50 in Algorithm 1) and upon success, it updates j to the new successfully allocated micro-slot (see line 48). The next micro-slot, at which to attempt transmission, will be $k + \lceil (j-k)/2 \rceil$. The search finishes when $\lfloor (j-k) \rfloor = 0$.

7.4. Performance Evaluation

In what follows, we evaluate the performance of DLMAC and BinDLMAC by conducting extensive simulations over different single- and multi-hop mm-Wave network scenarios. We compare our proposals with MDMAC [64], a recent link scheduling protocol for 60 GHz networks. Specifically, we measure the aggregate network throughput, when stations operate with the proposed schemes and respectively with different MDMAC versions[2], and transmit frames with varying payload sizes, under both homogeneous and heterogeneous data rates.

[2]By design, MDMAC works with a fixed slot size, optimized for a single payload. For a fair comparison, we examine the protocol's behavior with different slot sizes. We further discuss MDMAC's operation in Section 7.2.

For the evaluation, we implement the aforementioned schemes in a Matlab-based, event-driven simulator. We use the signal propagation model given in [79], with the following parameters: oxygen absorption coefficient $\alpha = 0.02$dB/m, carrier wavelength $\lambda = 5$mm, and transmit power $P = 10$dBm. In all the simulations, we configure DLMAC with $p_{\text{red}} = 0.2$ and $p_{\min} = 0.01$.

All protocols comply with the inter-frame and control message durations specified by the IEEE 802.11ad standard, as given in Table 7.1. We assume that the link data rates remain constant during simulation runtime. This assumption is supported by experimental results we obtained in our testbed, using a Dell 6430u laptop and a D5000 wireless docking system equipped with 60 GHz transceivers. These experiments confirm that MCS selection is consistent over 7-minutes tests, as illustrated in Fig. 7.5.

For all simulations, we give averages and 95% confidence intervals for the aggregate throughput, over 50 runs.

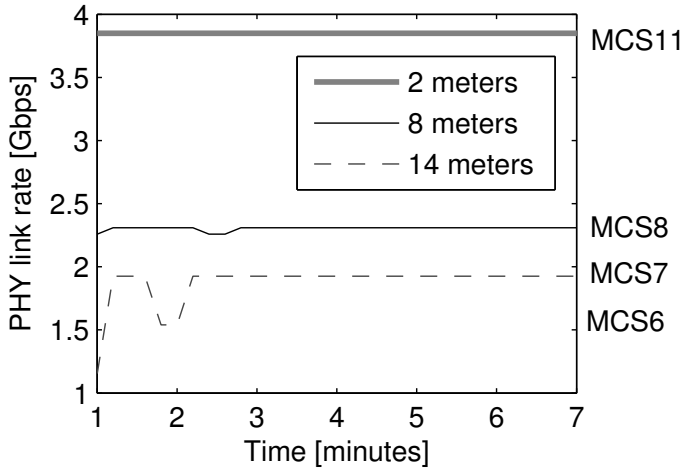


Figure 7.5: Experimental testbed results for the MCS selected by a laptop transmitting to a wireless docking station over the 60 GHz band for TX–RX distances of 2, 8, and 14 meters.

Table 7.1: IEEE 802.11ad [31] timing parameters.

Parameters	Values
aRTSTime	8.19μs
aCTSTime	8.19μs
aACKTime	6.45μs
aSIFSTime	3μs
aCTSTimeoutTime	15μs

7.4.1. Star and random topologies

We first consider two single-hop topologies with ten stations. In the first scenario, nodes transmit to the same AP (star topology), while in the second each station transmits to a randomly selected neighbor (random topology). All stations operate under saturation conditions (i.e., always have packets queued for transmission). Therefore, stations aim to perform multiple allocations within the same schedule. However, a node only attempts to find a new allocation once it successfully completed a packet exchange with an intended receiver. A station will refrain from allocating more transmissions within the same schedule once, to its knowledge, insufficient idle time remains to accommodate other stations.

We investigate scenarios where all stations transmit at a fixed data rate (1.925 Gbps), and respectively where each link operates with a randomly selected bit rate, ranging from 385 Mbps to 4.62 Gbps, corresponding to the 12 single carrier modulation and coding schemes (MCSs) defined by the IEEE 802.11ad standard [31]. We examine the performance of the protocols for different payload sizes $F = \{1.5, 3, 6, 12, 24\}$KB.

7.4.1.1. Star topology, homogeneous data rates

First we evaluate the throughput attained by DLMAC and BinDLMAC under homogeneous link conditions for different payload sizes and compare it to the performance of MDMAC configured with different slot sizes. We depict the results in Fig. 7.6. In line with our intuition, slotted channel access operating with a fixed slot size only works well when the payload fits the slot size perfectly. More specifically, *(i)* a small slot size leads to packet fragmentation, which may require multiple slots for a single transmission and thus incurs additional overhead (e.g. MDMAC-20μs, $F \geq$ 3KB); *(ii)* when the slot size is large, a fraction of the slot remains idle, which reduces protocol efficiency and thus the overall throughput (e.g. MDMAC-160μs, $\forall F$).

In contrast, the aggregate throughput of DLMAC increases monotonically with the payload size and approaches the maximum achievable value in all scenarios. This is due to the fact that DLMAC is inherently more flexible and assigns allocations dynamically.

We note however, that DLMAC attains lower throughput than MDMAC, when the payload exactly matches the slot size. For instance, a $F = 1.5$KB payload requires 19μs for transmission, leaving only 1μs idle time if the slot size is 20μs (MDMAC-20μs). This observation motivates the design of our BinDLMAC refinement, which seeks to further reduce the inter-transmission idle periods experienced by DLMAC.

Through the micro-slot binary search procedure, BinDLMAC successfully clusters transmissions, which leads to nearly optimal throughput performance. As seen in Fig. 7.6, by this procedure BinDLMAC achieves up to 25% more throughput than DLMAC and outperforms or performs very close to MDMAC in most settings.

To give further insight into the observed throughput difference, in Fig. 7.7 we plot the distribution of the inter-transmission idle time when DLMAC and BinDLMAC are used with 1.5 and

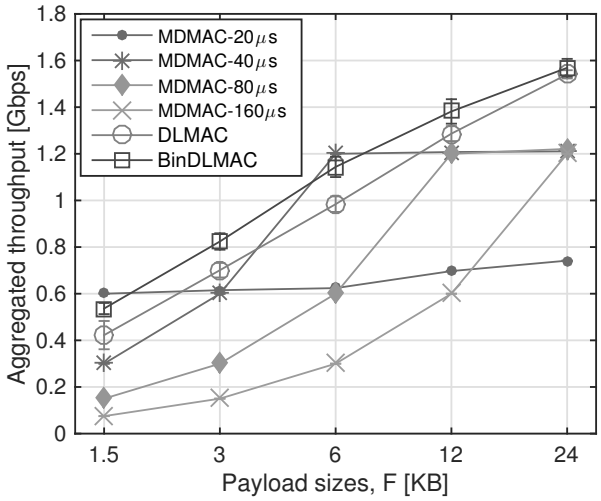
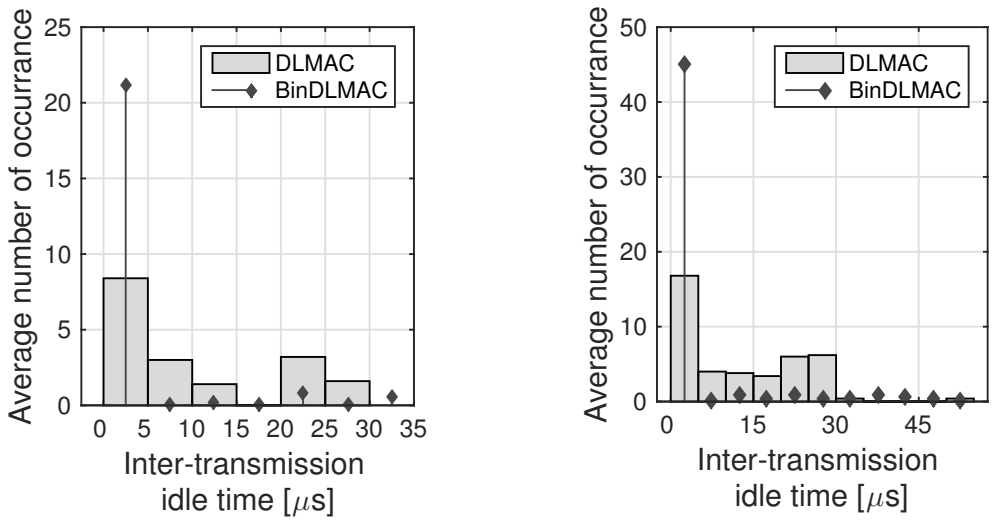


Figure 7.6: Throughput comparison between the proposed schemes (DLMAC and BinDLMAC) and slotted channel MDMAC with different slot sizes, for a star topology with $N = 10$ stations transmitting at 1.925Gbps.

6KB payloads. We observe that BinDLMAC does not eliminate large idle times completely, since the probabilistic probing we implement may create inter-cluster gaps. Despite this, BinDLMAC almost triples the number of very short idle intervals (0-5μs), while reducing the number of larger ones, which translates into the throughput gains illustrated in Fig. 7.6.



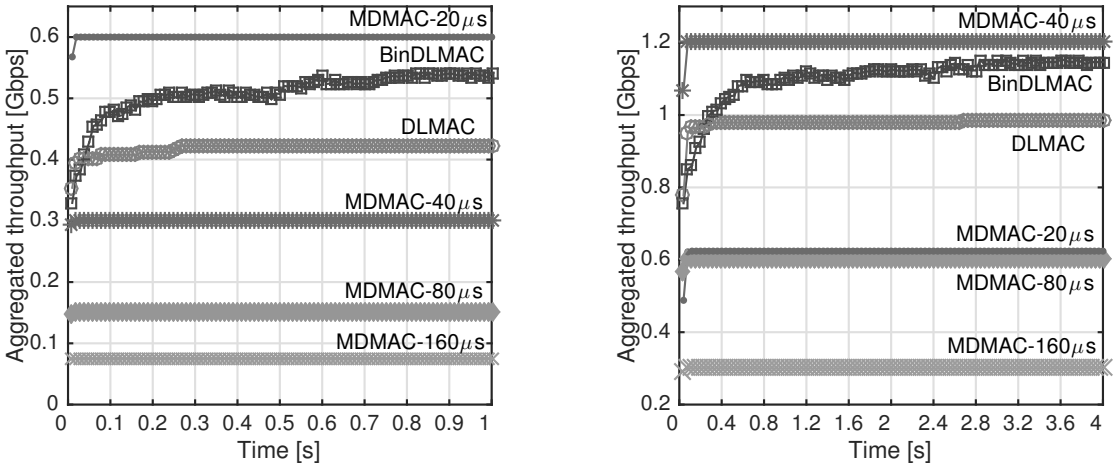
(a) Payload size = 1.5KB.

(b) Payload size = 6KB.

Figure 7.7: Inter-transmission idle time distributions for DLMAC and BinDLMAC, for 1.5KB (left) and 6KB (right) payloads.

We also examine DLMAC and BinDLMAC's convergence properties in the above scenario. For this purpose, in Fig. 7.8 we show the evolution of the aggregate throughput for both our

approaches and MDMAC. By design, MDMAC stabilizes quickly as all slots are allocated for transmission. In contrast, DLMAC takes slightly longer to converge due to the probing procedure employed to decrease the inter-transmission idle time. Since BinDLMAC further reduces these and improves channel utilization, it requires additional time to converge to a conflict free allocation. Nevertheless, we observe from Fig. 7.8 that both approaches settle in less than 1 (and respectively 4) second(s) when the payload size is 1.5KB (and respectively 6KB), which we consider acceptable for practical scenarios.



(a) Payload size = 1.5KB.

(b) Payload size = 6KB.

Figure 7.8: Evolution of aggregated throughput in a star topology ($N = 10$ nodes) for DLMAC and BinDLMAC, as well as MDMAC variants for comparison.

7.4.1.2. Star topology, heterogeneous data rates

Next, we demonstrate that our proposals achieve further performance gains over fixed slot size scheduling mechanisms when links operate with different data rates. We use the same star topology with $N = 10$ transmitters but with link data rates for the different stations chosen randomly from the set of 12 MCSs defined by the standard.

We illustrate the results in Fig. 7.9, where we plot the aggregate throughput of DLMAC, BinDLMAC, and the different MDMAC variants, as we vary the payload size. Observe that in this case, computing an optimal slot size that accommodates a frame perfectly is no longer feasible. As a consequence, all MDMAC variants perform poorly, as the payload size exceeds 1.5KB. In contrast, by employing unslotted channel access and allocating air time adaptively, DLMAC's performance is superior – our approach allocates transmission time individually, depending on both payload size and link rate; this overcomes underutilization of longer slots, as well as the increased overhead associated with short fixed slots. In addition, by reducing the duration of inter-transmission idle time, BinDLMAC further improves network throughput. Specifically, BinDLMAC achieves up to 100% more throughput than MDMAC-20μs and up to five times the

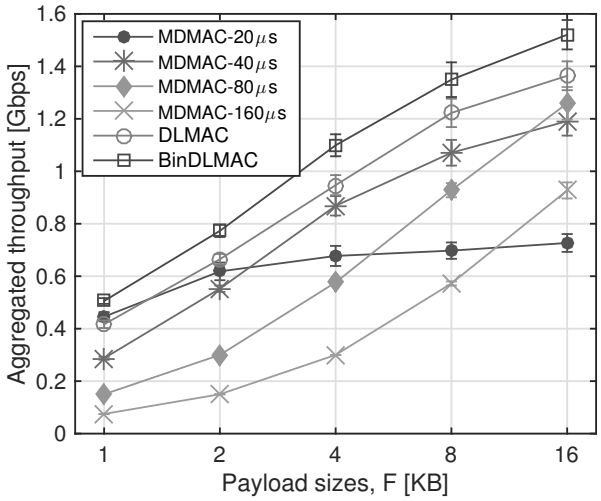


Figure 7.9: Throughput comparison between the proposed schemes DLMAC and BinDLMAC as well as the slotted channel MDMAC variants for different payload sizes, for a star topology with $N = 10$ stations with different data rates.

performance of MDMAC-160μs.

To verify that the observed performance gains are due to MDMAC experiencing high overhead (small slots) or leaving unnecessarily long idle periods within (long) slots, in Fig. 7.10 we show the percentage of air time for payload transmission (black), overhead (gray), and idle time (white), for both our schemes and the MDMAC variants. Indeed, DLMAC and BinDLMAC consistently utilize a higher fraction of time for payload transmission, while overhead decreases with payload size. In addition, idle time is reduced and protocol efficiency is further enhanced through our micro-slot binary search procedure.

We conclude that no unique slot size exists, such that the performance of slotted access schemes is maximized in all circumstances. By performing adaptive channel time allocation and clustering transmissions, DLMAC and BinDLMAC achieve superior throughput performance and substantially outperform the recently proposed MDMAC scheme.

7.4.1.3. Random topology

Next, we examine a scenario where transmitters do not share the same receiver. More specifically, we consider a 60 GHz network with $N = 10$ stations, where each node chooses a destination randomly, in both homogeneous and heterogeneous link rate scenarios. We demonstrate that in such topologies, the aggregate throughput gains of DLMAC and BinDLMAC over MDMAC variants are even higher. To this end, we plot again the network throughput as a function of the payload size when links operate with the same data rate (Fig. 7.11a) and for randomly chosen data rates among the set of allowed MCSs (Fig. 7.11b), respectively.

From Fig. 7.11 we conclude that in the random topologies evaluated, DLMAC achieves up to

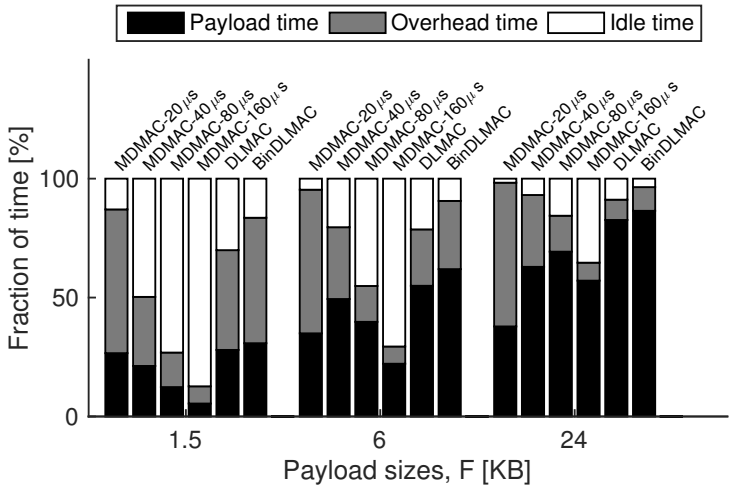
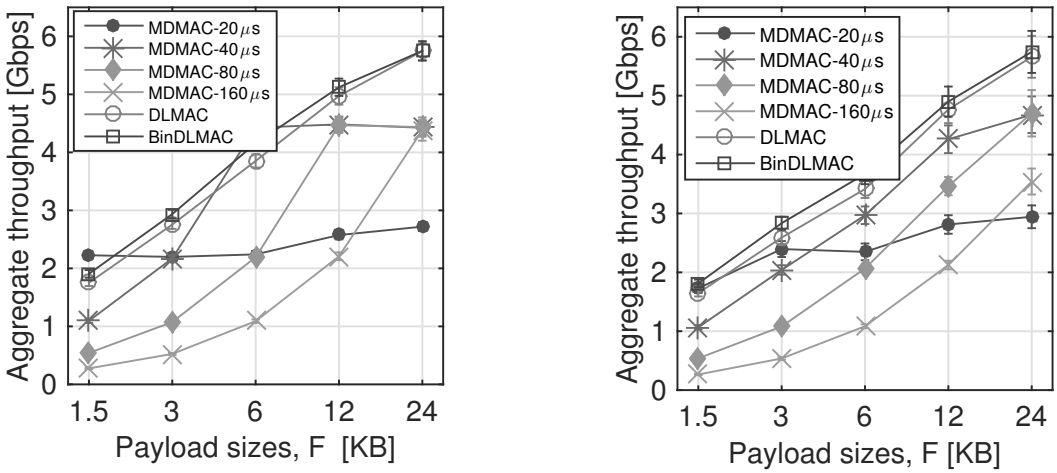


Figure 7.10: Fraction of time spent for payload transmission, packet overhead, as well as idle time, for DLMAC, BinDLMAC and the MDMAC variants for a star topology with $N = 10$ stations for different payload sizes and heterogeneous link rates.

8 times higher throughput than MDMAC. The BinDLMAC refinement succeeds in better packing transmissions, which results in further throughput improvements of up to 10% above DLMAC.



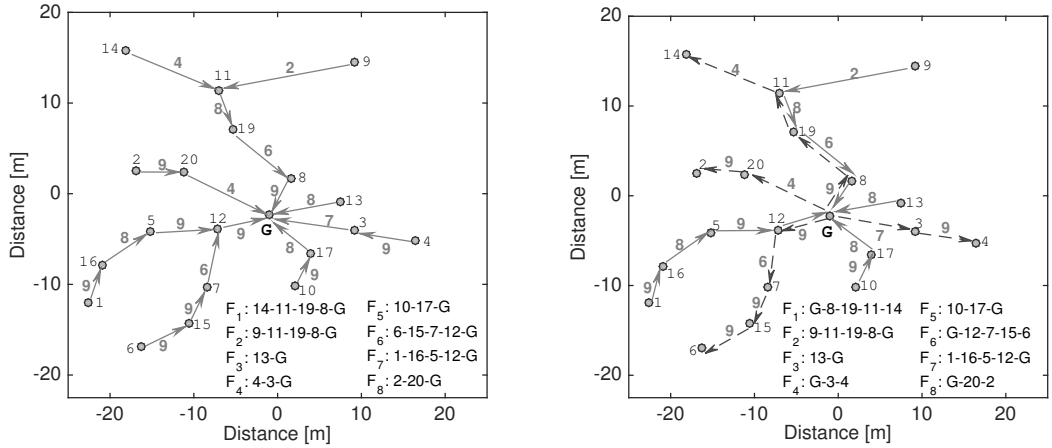
(a) Homogeneous link rate.

(b) Heterogeneous link rate.

Figure 7.11: Throughput comparison between the proposed schemes DLMAC and BinDLMAC as well as the MDMAC variants for a random single-hop topology with $N = 10$ stations with data rates of 1.925 Gbps (left) and rates ranging between 385Mbps and 4.62Gbps (right) for different payload sizes.

7.4.2. Multi-hop topologies

In what follows, we evaluate the performance of the proposed protocols in more complex network scenarios. Specifically, we consider a multi-hop network topology with 20 stations distributed across a 50mx50m area, as depicted in Fig. 7.12. We compute the corresponding data rate of each link based on the distance between communicating pairs and the propagation model specified by Zhu *et al.* [79], as described above. The antenna sector width of a station is 13° and the maximum distance between two communicating nodes is 23m (which follows the insights gained from our experiments in a real testbed). Although we do not capture the neighbor discovery phase, we note that this could be achieved through beam sweeping [31], or via omnidirectional transmissions in a lower frequency band, as suggested in [47].



(a) All flows towards gateway.

(b) Flows from and towards gateway.

Figure 7.12: Multi-hop topologies considered for evaluation, with links labeled with their corresponding MCS index (the corresponding data rate is shown in the Table 7.2) for a pure uplink scenario with all flows terminating at the gateway G (left) and a mixed uplink and downlink scenario (right).

Table 7.2: Mapping of MCS index to data rate as specified in [31].

MCS index	Data rate [Gbps]
2	0.7700
4	1.1550
6	1.5400
7	1.9250
8	2.3100
9	2.5025

In these simulations, we add further practical considerations, as well as complexity, by assuming flows operate with different payload sizes. Precisely, each flow randomly selects from a set of payloads $F = \{1.5, 3, 6, 12, 24\}$KB. We consider two distinct cases: *(i)* multiple flows

originate at different nodes and terminate at the gateway, as indicated by the labels in the bottom right corner of Fig. 7.12a; and *(ii)* several uplink and downlink flows coexist in the multi-hop topology, as indicated by the arrows and labels depicted in Fig. 7.12b.

In these scenarios, we measure the end-to-end throughput attained by all flows, and compute the average sum of the individual throughputs over 20 simulation runs, for DLMAC, BinDLMAC, and MDMAC with different slot sizes (between $20-160\mu$s). The results of these experiments are shown in Fig. 7.13, where we observe that also in these multi-hop topologies, as in the single-hop case, DLMAC and BinDLMAC attain substantially higher end-to-end throughput compared to MDMAC.

We conclude that, in multi-hop topologies with heterogeneous link rates and frame sizes, by employing unslotted channel access and clustering transmissions, BinDLMAC achieves between 20% and 160% throughput gains over MDMAC.

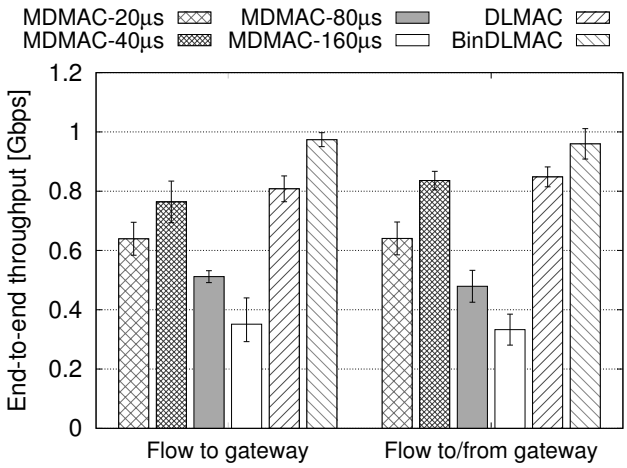


Figure 7.13: Comparison of the average sum of end-to-end throughputs attained by the flows shown in Fig. 7.12, when the stations operate with the proposed schemes and MDMAC variants.

7.5. Conclusions

Scheduling solutions for wireless networks mainly target scenarios in which carrier sensing is feasible. However, due to directional transmission used (to cope with critical path loss so as to achieve high throughput) in mm-Wave communications, nodes cannot rely on carrier sensing to assess the channel status. In this article, we tackled efficient scheduling for mm-Wave networks by applying a decentralized learning approach. In contrast to earlier works, we considered heterogeneous conditions in terms of link data rates and traffic demand across the network. By adopting a quasi-unslotted approach and finding allocations that result in successful transmissions while, at the same time, packing transmissions together to increase efficiency, the proposed protocols achieve 1.6 times the end-to-end throughput of existing approaches in heterogeneous multi-hop topologies, and even higher gains in single-hop scenarios. Moreover, our proposals do not re-

quire probabilistically resetting the protocol state to accommodate new transmissions, but instead ensure there is sufficient slack in the schedule to capture network dynamics. The proposed protocols neither require tight synchronization nor information exchange to build and maintain the schedule.

Chapter 8

Summary

This thesis explores the challenges in designing scheduling and medium access control schemes for wireless networks. In particular, we focus on problem arises due to the heterogeneity of the users due to the variability of wireless channels.

First of all, we identify that an optimal design for multicasting requires a precise tradeoff between multicast gain (i.e., the aim to transmit to as many users as possible) and multiuser diversity gain (i.e., the aim to achieve high throughput). Hence, we explicitly take into account two system parameters: *(i)* the channel quality, and *(ii)* the amount of data received at the users. This consideration further leads to another interesting tradeoff, which is completion time minimization versus throughput maximization at a given state for the decision made. The analysis of the optimal dynamic programming scheme reveals that throughput maximization is beneficial if all users still need to receive considerably large amounts of data from the base station. Otherwise, the lagging users are to be prioritized over the users that progress faster. Due to the complexity of dynamic programming, we design two simple and practical heuristics. Extensive simulations show that they perform close to the optimal solution. In more realistic scenarios with multipath Rayleigh fading and limited feedback message, they also achieve high gain as compared to the simple broadcast and opportunistic schemes.

Leveraging beamforming, we can improve the rate limitation of multicast scheduling when some users experience bad channel conditions. We first formulate the problem as a dynamic programming problem to obtain the optimal solution. Due to its complex nature, we design a heuristic algorithm that allows us to capture the characteristics of the solution in terms of selecting the optimal group of users to select. The evaluation of our heuristic algorithm in a discrete channel shows that it performs close to the optimal solution. The algorithm is additionally tested in a scenario with multipath Rayleigh fading as well as imperfect channel state information. Results show that our proposed scheme improves the completion time by up to 46.14% with respect to the benchmarked schemes (i.e., greedy and broadcast schemes) that only optimize for throughput and broadcast gain.

Next, we address the deafness problem in mm-Wave communications caused by directional

beamforming is used. Our research reveals that the deafness problem leads to excessive back-off and channel under-utilization. To mitigate this problem, we leverage omnidirectional transmissions at a lower frequency band. This allows us to dedicate the extremely high throughput mm-Wave band exclusively for data transmission. Results show that this dual-band approach solves the deafness problem, prevents fairness impairments and achieves much higher throughput.

Last but not least, we investigate the impact of the deafness problem for scheduling coordination in 60 GHz mesh networks. A self-organized scheduling mechanism is vital in these networks because carrier sensing is unreliable and central distribution of global messages is either difficult or infeasible. In contrast to previous work, we consider the heterogeneity of the network and propose a decentralized learning mechanism that enables scheduling independent of carrier sensing and global information.

8.1. Future work

In a nutshell, this thesis shows the important considerations to be taken into account when designing a scheduler as well as the MAC layer protocol for heterogeneous mobile networks. This thesis has explored *(i)* the advantages of multicasting techniques and *(ii)* the underlying problems and their potential solutions in mm-Wave communications. One interesting future topic of research is to investigate the feasibilities and challenges of multicasting in mm-Wave communications. Since the omni-directional transmission of mm-Wave transmission is weak, the multicasting capability in the mm-Wave band may be significantly lower than in the lower frequency bands (i.e., 2.4 or 5 GHz). In what follows, we present a more specific future work for each chapter in this thesis.

We evaluate our work on the optimization for finite horizon multicasting in Chapter 3 and Chapter 4 in extensive and diverse simulation scenarios. Since the simulation results of the algorithm are promising, experimental implementation and evaluation are valuable extensions to this work. Further, an actual implementation would also potentially reveal further challenges and network aspects that do not become evident in simulation.

Chapter 6 and Chapter 7 focus on solving the critical deafness problem faced by the highly attenuated signal in the millimeter-wave networks. In Chapter 6, we solve this problem by exploiting the multiple interfaces existing in mobile devices. We take advantage of the omni-directional capability of the legacy frequency band (i.e., 2.4 or 5 GHz) to exchange control messages between nodes. This ensures that the neighboring nodes are aware of ongoing transmissions. As a result, this solves the deafness problem faced by the directional link of mm-Wave communications. Despite the encouraging performance gains in terms of fairness and throughput, the information hand-over among frequency bands and the impact on traffic in the legacy band remains an open research issue. It is also worth to further investigate the additional implementation complexity and overhead needed to control the switching between the frequency bands.

Our proposal in Chapter 7.1 focuses on the design of a scheduler for a 60 GHz mesh network without modifying the standardized frame structure. We also avoid the above dependency on the legacy frequency band. Rather than listening to ongoing communications, a transmitter finds its allocation by learning from past communication events. While our evaluation results show that this scheduler works well in a static and quasi-static network environments, it would be interesting to explore its performance in scenarios with dynamic traffic as well as non-backlogged scenarios.

References

[1] R.O. Afolabi, A. Dadlani, and Kiseon Kim. Multicast Scheduling and Resource Allocation Algorithms for OFDMA-Based Systems: A Survey. *IEEE Communications Surveys Tutorials*, 15(1):240–254, Jan 2013.

[2] A. Arora, M. Krunz, and A. Muqattash. Directional Medium Access Protocol (DMAP) with Power Control for Wireless Ad Hoc Hetworks. In *IEEE GLOBECOM*, Nov 2004.

[3] E. Aryafar, M. Khojastepour, K. Sundaresan, S. Rangarajan, and E. Knightly. ADAM: An Adaptive Beamforming System for Multicasting in Wireless LANs. In *IEEE INFOCOM*, Mar 2012.

[4] A. Asadi and V. Mancuso. A Survey on Opportunistic Scheduling in Wireless Communications. *IEEE Communications Surveys Tutorials*, 15(4):1671–1688, Apr 2013.

[5] A. Asadi and V. Mancuso. On the Compound Impact of Opportunistic Scheduling and D2D Communications in Cellular Networks. In *ACM MSWiM*, Aug 2013.

[6] J. Barcelo, B. Bellalta, C. Cano, and M. Oliver. Learning-BEB: Avoiding Collisions in WLANs. *Carrier Sense Multiple Access with Enhanced Collision Avoidance*, page 23, 2009.

[7] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. I*. Athena Scientific, 3rd edition, 2005.

[8] G. Bianchi. Performance Analysis of the IEEE 802.11 Distributed Coordination Function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547, Mar 2000.

[9] C. Cano, D. Malone, B. Bellalta, and J. Barceló. On the Improvement of Receiver-Initiated MAC Protocols for WSNs by Applying Scheduling. In *IEEE WoWMoM*, Jun 2013.

[10] P. Cerwall (ed). Ericsson Mobility Report. `http://www.ericsson.com/mobility-report`, June 2015.

[11] E. Chai, K.G. Shin, S.-J. Lee, J. Lee, and R. Etkin. Defeating Heterogeneity in Wireless Multicast Networks. In *IEEE INFOCOM*, Apr 2013.

[12] K. Chandra, R.V. Prasad, I.GMM. Niemegeers, and A.R. Biswas. Adaptive Beamwidth Selection for Contention Based Access Periods in Millimeter Wave WLANs. In *IEEE CCNC*, Jan 2014.

[13] T.-H. Chang, Z.-Q. Luo, and C.-Y. Chi. Approximation Bounds for Semidefinite Relaxation of Max-Min-Fair Multicast Transmit Beamforming Problem. *IEEE Transactions on Signal Processing*, 56(8):3932–3943, Jul 2008.

[14] Q. Chen, J. Tang, D.T.C. Wong, X. Peng, and Y. Zhang. Directional Cooperative MAC Protocol Sesign and Performance Analysis for IEEE 802.11 ad WLANs. *IEEE Transactions on Vehicular Technology*, 62(6):2667–2677, Jul 2013.

[15] R.R. Choudhury, X. Yang, R. Ramanathan, and N.H. Vaidya. On Designing MAC Protocols for Wireless Networks Using Directional Antennas. *IEEE Transactions on Mobile Computing*, 5(5):477–491, May 2006.

[16] Rui del Negro. Bitrate & GOP calculator, 2014.

[17] L.G. Didier. MPEG: A Video Compression Standard for Multimedia Applications. *Communications of the ACM - Special issue on digital multimedia systems*, 34(4):46–58, Apr 1991.

[18] Y. Du, E. Aryafar, J. Camp, and M. Chiang. iBeam: Intelligent Client-side Multi-user Beamforming in Wireless Networks. In *IEEE INFOCOM*, Apr 2014.

[19] M. Fang, D. Malone, K.R. Duffy, and D.J. Leith. Decentralised Learning MACs for Collision-free Access in WLANs. *Wireless Networks*, 19(1):83–98, 2013.

[20] W.Y. Ge, J.S. Zhang, and S. Shen. A Cross-Layer Design Approach to Multicast in Wireless Networks. *IEEE Transactions on Wireless Communications*, 6(3):1063–1071, Mar 2007.

[21] D. Gong, Y.Y Yang, and H.W. Li. Link-Layer Multicast in Smart Antenna Based 802.11n Wireless LANs. In *IEEE MASS*, Oct 2013.

[22] M.X. Gong, D. Akhmetov, R. Want, and Shiwen Mao. Multi-User Operation in mmWave Wireless Networks. In *IEEE ICC*, Jun 2011.

[23] M.X. Gong, R. Stacey, D. Akhmetov, and Shiwen Mao. A Directional CSMA/CA Protocol for mmWave Wireless PANs. In *IEEE WCNC*, Apr 2010.

[24] D.H. Han, J.W. Jwa, and H.I. Kim. A Dual-Channel MAC Protocol Using Directional Antennas in Location Aware Ad Hoc Networks. In *ICCSA*, number 3983 in Lecture Notes in Computer Science, pages 594–602. Springer Berlin Heidelberg, Jan 2006.

[25] Y. He, J. Sun, X. Ma, A.V. Vasilakos, R. Yuan, and W. Gong. Semi-random Backoff: Towards Resource Reservation for Channel Access in Wireless LANs. *IEEE/ACM Transaction on Networking*, 21(1):204–217, Feb 2013.

[26] Q.-D. Ho and L.-N. Tho. Adaptive Opportunistic Multicast Scheduling Over Next-Generation Wireless Networks. *Wireless Personal Communication*, 63(2):483–500, Mar 2012.

[27] K. Hosoya, N. Prasad, K. Ramachandran, N. Orihashi, S. Kishimoto, S. Rangarajan, and K. Maruhashi. Multiple Sector ID Capture (MIDC): A Novel Beamforming Technique for 60-GHz Band Multi-Gbps WLAN/PAN Systems. *IEEE Transaction on Antennas and Propagation*, 63(1):81–96, Jan 2015.

[28] J. Huang, F. Qian, A. Gerber, Z.M. Mao, S. Sen, and O. Spatscheck. A Close Examination of Performance and Power Characteristics of 4G LTE Networks. In *ACM MobiSys*, Jun 2012.

[29] S.-M. Huang, J.-N. Hwang, and Y.-C. Chen. Reducing Feedback Load of Opportunistic Multicast Scheduling over Wireless Systems. *IEEE Communications Letters*, 14(12):1179–1181, Dec 2010.

[30] IEEE. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band. *IEEE Std 802.11ad-2012*, pages 1–628, Dec 2012.

[31] IEEE. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band . *IEEE 802.11ad Std.*, pages 1–634, Mar 2014.

[32] R. Jain, A. Durresi, and G. Babic. Throughput Fairness Index: An Explanation. Technical report, Department of CIS, The Ohio State University, 1999.

[33] Y.-B. Ko, V. Shankarkumar, and N.F. Vaidya. Medium Access Control Protocols Using Directional Antennas in Ad Hoc Networks. In *IEEE INFOCOM*, Mar 2000.

[34] T. Korakis, G. Jakllari, and L. Tassiulas. A MAC Protocol for Full Exploitation of Directional Antennas in Ad-hoc Wireless Networks. In *ACM MobiHoc*, Jun 2003.

[35] U.C. Kozat. On the Throughput Capacity of Opportunistic Multicasting with Erasure Codes. In *IEEE INFOCOM*, Apr 2008.

[36] R. Laufer, T. Salonidis, H. Lundgren, and P. Le Guyadec. XPRESS: A Cross-layer Backpressure Architecture for Wireless Multi-hop Networks. In *ACM MobiCom*, Sep 2011.

[37] J. Lee and Jean C.W. Design and Analysis of an Asynchronous Zero Collision MAC Protocol. *CoRR*, abs/0806.3542, 2008.

[38] T.-P. Low, M.-O. Pun, and C.-C. Jay Kuo. Optimized Opportunistic Multicast Scheduling over Cellular Networks. In *IEEE Globecom*, Dec 2008.

[39] T.-P. Low, M.-O. Pun, Y.-W. Peter Hong, and C.-C. Jay Kuo. Optimized Opportunistic Multicast Scheduling (OMS) over Heterogeneous Cellular Networks. In *IEEE ICASSP*, Apr 2009.

[40] T.-P. Low, M.-O. Pun, Y.-W. Peter Hong, and C.-C. Jay Kuo. Multi-Antenna Multicasting with Opportunistic Multicast Scheduling and Space-Time Transmission. In *IEEE GLOBECOM*, Dec 2010.

[41] T.-P. Low, M.-O. Pun, Y.-W. Peter Hong, and C.-C. Jay Kuo. Optimized opportunistic multicast scheduling (OMS) over wireless cellular networks. *IEEE Transactions on Wireless Communications*, 9(2):791 –801, Feb 2010.

[42] C. Mehlführer, M. Wrulich, J.C. Ikuno, D. Bosanska, and M. Rupp. Simulating the Long Term Evolution Physical Layer. In *IEEE EUSIPCO*, volume 27, page 124, 2009.

[43] S. Misra and M. Khatua. Semi-Distributed Backoff: Collision-Aware Migration from Random to Deterministic Backoff. *IEEE Transactions on Mobile Computing*, 14(5):1071–1084, May 2015.

[44] R. Mudumbai, S. Singh, and U. Madhow. Medium Access Control for 60 GHz Outdoor Mesh Networks with Highly Directional Links. In *IEEE INFOCOM*, Apr 2009.

[45] A. Nasipuri, S. Ye, J. You, and R.E. Hiromoto. A MAC Protocol for Mobile Ad Hoc Networks using Directional Antennas. In *IEEE WCNC*, Apr 2000.

[46] T. Nitsche, C. Cordeiro, A.B. Flores, E.W. Knightly, E. Perahia, and J.C. Widmer. IEEE 802.11ad: Directional 60 GHz Communication for Multi-Gigabit-per-second Wi-Fi [Invited Paper]. *IEEE Communications Magazine*, 52(12):132–141, Dec 2014.

[47] T Nitsche, A.B Flores, E.W. Knightly, and J.C Widmer. Steering with Eyes Closed: mm-Wave Beam Steering without In-Band Measurement. In *IEEE INFOCOM*, Apr 2015.

[48] H. Park, Y. Kim, T. Song, and S. Pack. Multi-band Directional Neighbor Discovery in Self-Organized mmWave Ad-hoc Networks. *IEEE Transactions on Vehicular Technology*, PP(99):1–1, Jun 2014.

[49] K.G. Praveen and E.G. Hesham. Opportunistic Multicasting. In *IEEE ASILOMAR*, Nov 2004.

[50] K.G. Praveen and E.G. Hesham. On the Throughput-delay Tradeoff in Cellular Multicast. In *IEEE IWCMC*, Jun 2005.

[51] Jian Q., L.X. Cai, X. Shen, and Jon W. Mark. Enabling Multi-Hop Concurrent Transmissions in 60 GHz Wireless Personal Area Networks. *IEEE Transactions on Wireless Communications.*, 10(11):3824–3833, Nov 2011.

[52] J. Qiao, X. Shen, J.W. Mark, Z. Shi, and N. Mohammadizadeh. MAC-layer Integration of Multiple Radio Bands in Indoor Millimeter Wave Networks. In *IEEE WCNC*, Apr 2013.

[53] T.S. Rappaport, S. Shu, R. Mayzus, Z. Hang, Y. Azar, K. Wang, G.N. Wong, J.K. Schulz, M. Samimi, and F. Gutierrez. Millimeter Wave Mobile Communications for 5G Cellular: It Will Work! *IEEE Access*, 1:335–349, May 2013.

[54] ITURM Recommendation. A Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000. *International Telecommunication Union*, 1997.

[55] L. Sanabria-Russo, J. Barcelo, and B. Bellalta. A High Efficiency MAC Protocol for WLANs: Providing Fairness in Dense Scenarios. *Submitted to Transaction on Networking*, Dec 2014. arXiv: 1412.1395.

[56] S. Sen, Jie X., R. Ghosh, and R.R. Choudhury. Link Layer Multicasting with Amart Antennas: No Client Left Behind. In *IEEE ICNP*, Oct 2008.

[57] S. Sesia, I. Toufik, and M. Baker. *LTE: the UMTS Long Term Evolution*. Wiley Online Library, 2009.

[58] W.L. Shen, K.C.-J. Lin, S. Gollakota, and M.S. Chen. Rate Adaptation for 802.11 Multiuser MIMO Networks. *IEEE Transactions on Mobile Computing*, 13(1):35–47, Jan 2014.

[59] A. Shokrollahi. An Introduction to Low-density Parity-check Codes. In *Theoretical aspects of computer science*, pages 175–197. Springer, 2002.

[60] A. Shokrollahi. Raptor Codes. *IEEE Transactions on Information Theory*, 52(6):2551–2567, Jun 2006.

[61] N.D. Sidiropoulos, T.N. Davidson, and Zhi-Quan L. Transmit Beamforming for Physical-layer Multicasting. *IEEE Transactions on Signal Processing*, 54(6):2239–2251, Jun 2006.

[62] G.H. Sim, B. Rengarajan, and J. Widmer. Adaptive Modulation for Finite Horizon Multicasting of Erasure-coded Data. In *IEEE COMSNETS*, Jan 2013.

[63] G.H. Sim, J. Widmer, and B. Rengarajan. Opportunistic Finite Horizon Multicasting of Erasure-coded Data. *IEEE Transactions on Mobile Computing*, PP(99):1–1, Apr 2015.

[64] S. Singh, R. Mudumbai, and U. Madhow. Distributed Coordination with Deaf Neighbors: Efficient Medium Access for 60 GHz Mesh Networks. In *IEEE INFOCOM*, Mar 2010.

[65] S. Singh, R. Mudumbai, and U. Madhow. Interference Analysis for Highly Directional 60-GHz Mesh Networks: The Case for Rethinking Medium Access Control. *IEEE ACM Transaction on Networking*, 19(5):1513–1527, Oct 2011.

[66] S. Singh, F. Ziliotto, U. Madhow, E. Belding, and M. Rodwell. Blockage and Directivity in 60 GHz Wireless Personal Area Networks: from Cross-layer Model to Multihop MAC Design. *IEEE Journal on Selected Areas in Communications*, 27(8):1400–1413, Oct 2009.

[67] K. Sundaresan, K. Ramachandran, and S. Rangarajan. Optimal Beam Scheduling for Multicasting in Wireless Networks. In *ACM MobiCom*, Sep 2009.

[68] M. Takai, J. Martin, R. Bagrodia, and A. Ren. Directional Virtual Carrier Sensing for Directional Antennas in Mobile Ad Hoc Networks. In *ACM MobiHoc*, Jun 2002.

[69] X. Tie, K. Ramachandran, and R. Mahindra. On 60 GHz Wireless Link Performance in Indoor Environments. In *PAM*, Mar 2012.

[70] M.F. Tuysuz and H.A. Mantar. A Beacon-Based Collision-Free Channel Access Scheme for IEEE 802.11 WLANs. *Wireless Personal Communications*, 75(1):155–177, Aug 2013.

[71] E. Veshi, A. Kuehne, and A. Klein. Comparison of Different Multicast Strategies in Wireless Identically Distributed Channels. In *IEEE WCNC*, Apr 2013.

[72] Y. Wang, H.K. Garg, and M. Motani. Directional Medium Access Control for Ad Hoc Networks: A Cooperation-based Approach. In *IEEE ICON*, Dec 2013.

[73] Y. Wang, X. Wang, M. Li, and J. Qu. A Parity-Based Opportunistic Multicast Scheduling Scheme over Cellular Networks. In *IEEE DASC*, pages 565–568, Dec 2013.

[74] H. Wen and Yeung K.L. Optimal Opportunistic Multicast for Minimizing Broadcast Latency in Wireless Networks. In *IEEE ICC*, May 2010.

[75] H. Wen and K.L. Yeung. On Maximizing the Throughput of Opportunistic Multicast in Wireless Cellular Networks with Erasure Codes. In *IEEE ICC*, Jun 2011.

[76] S. Yong, P. Xia, and A. Valdes-Garcia. *60GHz Technology for Gbps WLAN and WPAN: from Theory to Practice*. John Wiley & Sons, 2011.

[77] W. Zame, Jie Xu, and M. van der Schaar. Winning the Lottery: Learning Perfect Coordination with Minimal Feedback. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):846–857, Oct 2013.

[78] H. Zhang, Y. Jiang, K. Sundaresan, S. Rangarajan, and B. Zhao. Wireless Multicast Scheduling with Switched Beamforming Antennas. *IEEE/ACM Transactions on Networking*, 20(5):1595–1607, Oct 2012.

[79] Y. Zhu, Z. Zhang, Z. Marzi, C. Nelson, U. Madhow, BY. Zhao, and H Zheng. Demystifying 60GHz Outdoor Picocells. In *ACM Mobicom*, Sep 2014.